UNIVERSITY OF MALAYA

The Leader in Research & Innovation

MQB7046 – MODELLING PUBLIC HEALTH DATA

CONTINUOUS ASSESSMENT 4

CHIN WEI HONG (22110451)

MOHAMAD FADZIL BIN ABD. RAHIM (22079894)

ZAW MYO HTET (22109084)

DEPARTMENT OF SOCIAL AND PREVENTIVE MEDICINE

FACULTY OF MEDICINE, UNIVERSITY MALAYA

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

MQB7046: MODELLING PUBLIC HEALTH DATA

Semester 2, Session 2023/2024

CONTINUOUS ASSESSMENT 4

1. INTRODUCTION

1.1 Background

Liver disease is a significant global health concern, contributing to substantial morbidity and mortality rates worldwide. The liver, being a vital organ, performs numerous essential functions, including detoxification, protein synthesis, and the production of biochemicals necessary for digestion. Consequently, liver diseases can severely impact an individual's health and quality of life. Early diagnosis and effective management of liver diseases are crucial for improving patient outcomes and reducing healthcare burdens.

Liver diseases encompass a wide range of conditions, including hepatitis, cirrhosis, and liver cancer. These conditions can be caused by various factors such as viral infections, alcohol consumption, obesity, and genetic predispositions. The biochemical markers included in the dataset, such as Total Bilirubin (TB), Direct Bilirubin (DB), Total Proteins (TP), Albumin (ALB), Alanine Aminotransferase (AST), Aspartate Aminotransferase (ALT), and Alkaline Phosphatase (ALP), are critical indicators of liver function and health. Abnormal levels of these markers can signal liver damage or dysfunction.

Liver disease in Malaysia, particularly viral hepatitis B and C (HBV, HCV) and non-alcoholic fatty liver disease (NAFLD), poses a significant health burden, with these conditions being the leading causes of liver-related deaths in the country. Studies have highlighted gaps in public knowledge regarding different types of hepatitis, transmission risks, and complications of HBV and HCV, as well as low awareness of NAFLD risk factors, screening tests, and complications (Mohamed et al., 2023).

Furthermore, research on psoriasis patients in Malaysia revealed that 0.8% of them had liver disease, with viral hepatitis, fatty liver, and liver cirrhosis being the most common conditions, emphasizing the intersection of liver diseases with other health issues (Lim et al., 2023). Additionally, investigations into genetic polymorphisms of pro-inflammatory cytokines like IL-1b did not show a significant association with HCV infection susceptibility among Malay male drug abusers, indicating the complex interplay of genetic factors in liver disease development (Lansayan et al., 2023). These findings underscore the importance of robust education efforts and comprehensive healthcare strategies to address liver diseases in Malaysia effectively (S. Z. Lim et al., 2022).

1.1.1 Motivation and Importance

The dataset under consideration contains records of 553 patients, with a mix of individuals diagnosed with and without liver disease. The primary objective is to predict the presence of liver disease based on various biochemical markers, clinical profiles, and sociodemographic information. This predictive task is of paramount importance as it can aid in the early detection of liver diseases, enabling timely intervention and treatment. The significance of this dataset lies in its potential to enhance our understanding of liver disease diagnostics. By leveraging machine learning algorithms, we can develop predictive models that accurately identify patients at risk of liver disease based on their biochemical and clinical profiles. Such models can serve as valuable tools for healthcare professionals, enabling them to make informed decisions regarding patient care and management.

1.2 Literature Review

1.2.1 Biochemical Markers in Liver Disease Diagnosis

Biochemical markers play a crucial role in diagnosing various liver diseases, with bilirubin, albumin, and liver enzymes (AST, ALT, ALP) are critical indicators of liver function and damage (Kim et al., 2002). Elevated levels of these markers are often associated with liver disease. For instance, research has shown that liver function tests, including ALP, ALT, AST, and TB levels, are essential in assessing liver disease severity and progression, with chronic patients exhibiting distinct biochemical profiles compared to carriers (Ikonnikova et al., 2022). Additionally, advancements in non-invasive diagnostic tools, such as serum biomarkers and imaging modalities, have been proposed to overcome the limitations of traditional methods like liver biopsy, offering safer, more accessible, and cost-effective alternatives for liver disease evaluation and monitoring (Al-Salih et al., 2021).

1.2.2 Sociodemographic Factors and Liver Disease

Research has shown that sociodemographic factors, such as age, gender, and ethnicity, influence the prevalence and progression of liver disease. Certain ethnic groups may have a higher genetic predisposition to liver disease. Neighborhood-level social determinants of health (SDOH), including affluence and disadvantage, significantly impact mortality, liver-related events, and cardiovascular disease in patients with steatotic liver disease (Chen et al., 2023). Additionally, aging is a key factor in liver cancer mortality, highlighting the need for tailored prevention strategies (Luo et al., 2022). Socioeconomic status also plays a crucial role; for example, public assistance recipients have higher mortality rates in cases of alcoholic liver cirrhosis (Kushibuchi et al., 2022). Disparities in pediatric liver disease are evident, with black and Hispanic populations disproportionately affected and non-alcoholic fatty liver disease (NAFLD) becoming a rapidly rising concern in children (Martin et al., 2020). In rural India, risk factors for chronic liver disease include low education, poor socioeconomic status, diabetes, high BMI, alcohol consumption, and tobacco use, underscoring the multifactorial nature of liver disease etiology (Banait et al., 2021).

1.2.3 Machine Learning in Liver Disease Prediction

Machine learning plays a crucial role in predicting and diagnosing liver diseases, offering significant advancements in healthcare. Various machine learning models such as Naive Bayes, Support Vector Machine, Logistic Regression, Decision Tree, Random

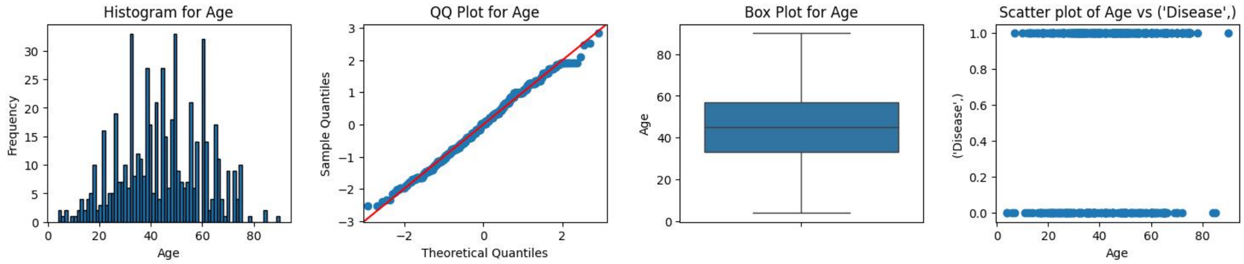
Forest Classifier, K-Nearest Neighbours Classifier, Kernel SVM, and XGBoost have been successfully implemented to predict chronic liver disease with high accuracy, as demonstrated in studies like those by Bhushan et al. (2023), Balaji et al. (2023), and Swetha et al. (2023). These models aid in early detection of liver diseases caused by factors like excessive alcohol use, drug use, and hepatitis, enabling timely intervention and treatment. Integrating artificial intelligence with healthcare through machine learning algorithms like Support Vector Machine, K-Nearest Neighbor, and Artificial Neural Network enhances the efficiency and accuracy of liver disease diagnosis, potentially reducing healthcare costs and improving patient outcomes (Nigatu et al., 2023).

1.3 Descriptive Statistics

The dataset comprises records of 553 patients, with various biochemical markers, clinical profiles, and sociodemographic information. The descriptive statistics for each feature are presented below:

1.3.1 Continous Data

Age: The age of the patients ranges from 4 to 90 years (M = 44.43, SD = 16.01). The distribution of age is approximately normal (Shapiro-Wilk W = 0.993, p = 0.012), indicating a slight deviation from normality.


Histogram for Age

QQ Plot for Age

Box Plot for Age

Scatter plot of Age vs ('Disease',)

Figure 1.1 Visual Summary of Age

Albumin and Globulin Ratio (AGR): The AGR values range from 0.3 to 2.8 (M = 0.95, SD = 0.32). The distribution is positively skewed (skewness = 1.00) and leptokurtic (kurtosis = 3.37), indicating a higher frequency of lower values. The Shapiro-Wilk test indicates a significant deviation from normality (W = 0.946, p < 0.001).

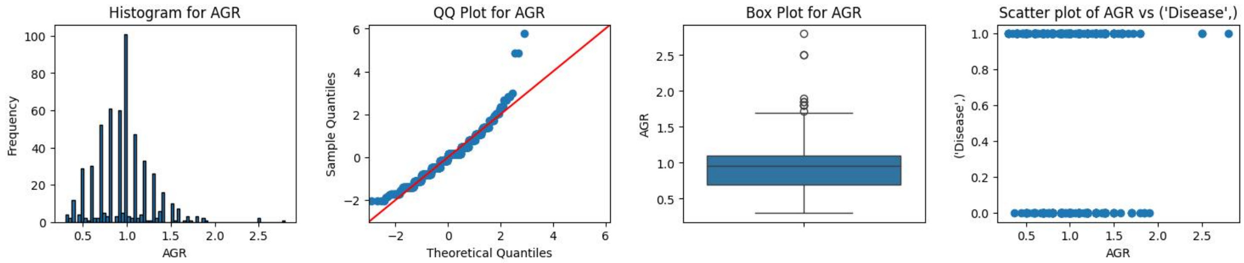


Figure 1.2 Visual Summary of Albumin and Globulin Ratio

Total Bilirubin (TB): The total bilirubin levels range from 0.4 to 42.8 (M = 3.13, SD = 5.41). The distribution is highly positively skewed (skewness = 3.46) and leptokurtic (kurtosis = 13.61), indicating a higher frequency of lower values. The Shapiro-Wilk test indicates a significant deviation from normality (W = 0.505, p < 0.001).

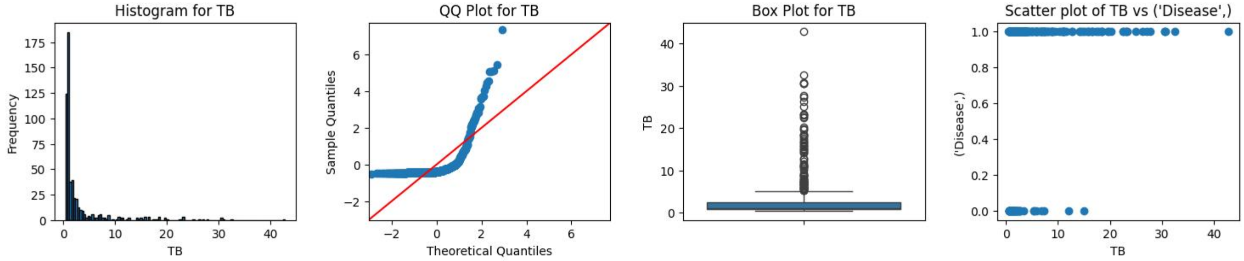


Figure 1.3 Visual Summary of Total Bilirubin

Direct Bilirubin (DB): The direct bilirubin levels range from 0.1 to 19.7 (M = 1.46, SD = 2.78). The distribution is highly positively skewed (skewness = 3.27) and leptokurtic (kurtosis = 11.86), indicating a higher frequency of lower values. The Shapiro-Wilk test indicates a significant deviation from normality (W = 0.525, p < 0.001).

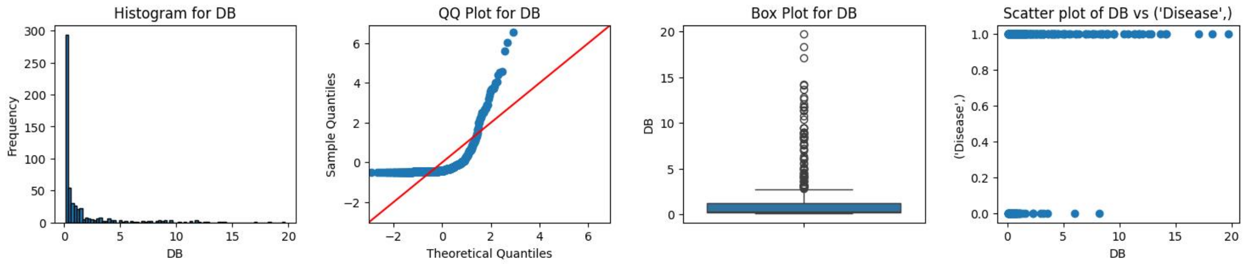


Figure 1.4 Visual Summary of Direct Bilirubin

Albumin (ALB): The albumin levels range from 0.9 to 5.5 (M = 3.15, SD = 0.79). The distribution is approximately normal (W = 0.992, p = 0.004), with a slight negative skewness (skewness = -0.10) and a slight platykurtic distribution (kurtosis = -0.44).

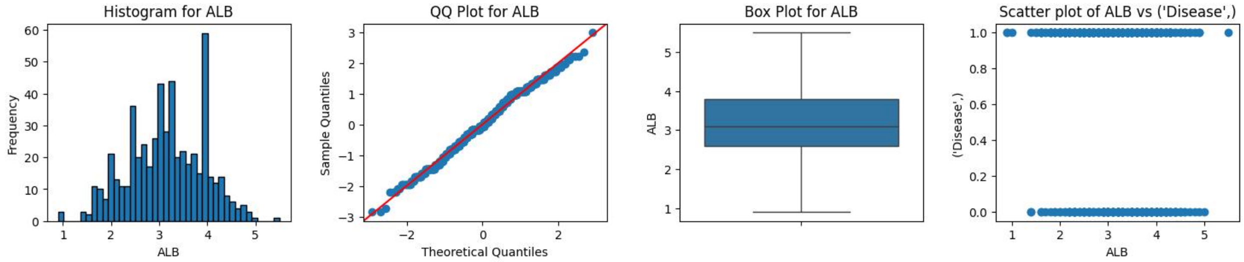


Figure 1.5 Visual Summary of Albumin

Total Proteins (TP): The total protein levels range from 2.7 to 9.6 (M = 6.48, SD = 1.08). The distribution is approximately normal (W = 0.991, p = 0.002), with a slight negative skewness (skewness = -0.33) and a near-normal kurtosis (kurtosis = 0.21).

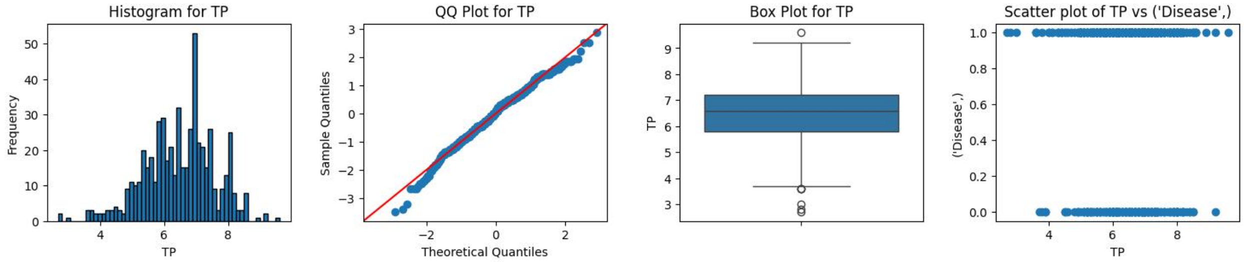


Figure 1.6 Visual Summary of Total Proteins

Alkaline Phosphatase (ALP): The alkaline phosphatase levels range from 63 to 2110 (M = 286.38, SD = 239.58). The distribution is highly positively skewed (skewness = 3.82) and leptokurtic (kurtosis = 18.37), indicating a higher frequency of lower values. The Shapiro-Wilk test indicates a significant deviation from normality (W = 0.580, p < 0.001).

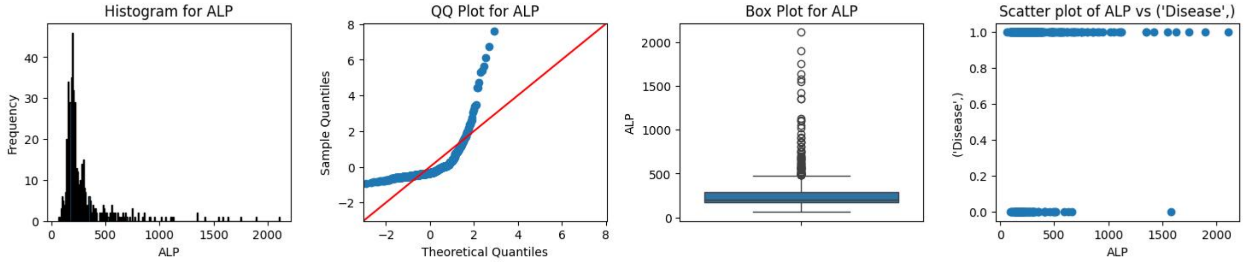


Figure 1.7 Visual Summary of Alkaline Phosphatase

Alanine Aminotransferase (ALT): The alanine aminotransferase levels range from 10 to 4929 (M = 111.81, SD = 295.69). The distribution is highly positively skewed (skewness = 10.32) and leptokurtic (kurtosis = 143.42), indicating a higher frequency of lower values. The Shapiro-Wilk test indicates a significant deviation from normality (W = 0.281, p < 0.001).

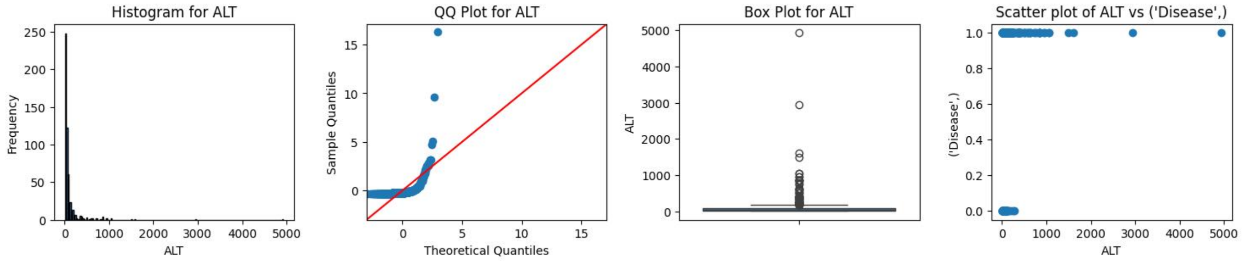


Figure 1.8 Visual Summary of Alanine Aminotransferase

Aspartate Aminotransferase (AST): The aspartate aminotransferase levels range from 10 to 2000 (M = 81.50, SD = 186.84). The distribution is highly positively skewed (skewness = 6.41) and leptokurtic (kurtosis = 48.02), indicating a higher frequency of lower values. The Shapiro-Wilk test indicates a significant deviation from normality (W = 0.327, p < 0.001).

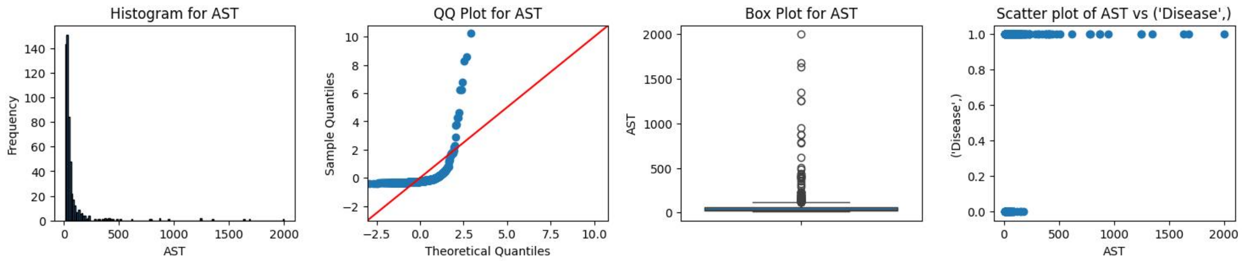


Figure 1.9 Visual Summary of Aspartate Aminotransferase

1.3.2 Categorical Data

Table 1.1 Descriptive Analaysis for Continous Data

Variables	Have Liver Disease	No liver disease	All	percentage (%)
Gender				
Female	87	50	137	24.77
Male	296	120	416	75.23
Ethnici				
Chinese	110	29	139	25.14
Indian	93	46	139	25.14
Malay	180	95	275	49.73
BMI				
Normal BMI	70	128	198	35.8
Overweight	313	42	355	64.2

Gender: The analysis shows that 87 females and 296 males have liver disease, while 50 females and 120 males do not have liver disease. The chi-square test indicates no significant association between gender and liver disease (χ2 = 2.49 , p = 0.11).

Ethnicity: The analysis reveals that 110 Chinese, 93 Indians, and 180 Malays have liver disease, while 29 Chinese, 46 Indians, and 95 Malays do not have liver disease. The chi-square test suggests a significant association between ethnicity and liver disease (χ2 = 8.60 , p = 0.01).

BMI: The analysis shows that 70 individuals with normal BMI and 313 overweight individuals have liver disease, while 128 individuals with normal BMI and 42 overweight individuals do not have liver disease. The chi-square test indicates a significant association between BMI and liver disease (χ2 = 164.06 , p < 0.001).

Disease: The target variable indicates the presence of liver disease, with 69.26% of the patients diagnosed with liver disease (M = 0.69, SD = 0.46). The distribution is negatively skewed (skewness = -0.83) and platykurtic (kurtosis = -1.30). The Shapiro-Wilk test indicates a significant deviation from normality (W = 0.580, p < 0.001).

1.4 Correlation Matrix

The correlation matrix provides insights into the relationships between various features in the dataset. The correlation coefficients range from -1 to 1, where values close to 1 indicate a strong positive correlation, values close to -1 indicate a strong negative correlation, and values around 0 indicate no correlation.

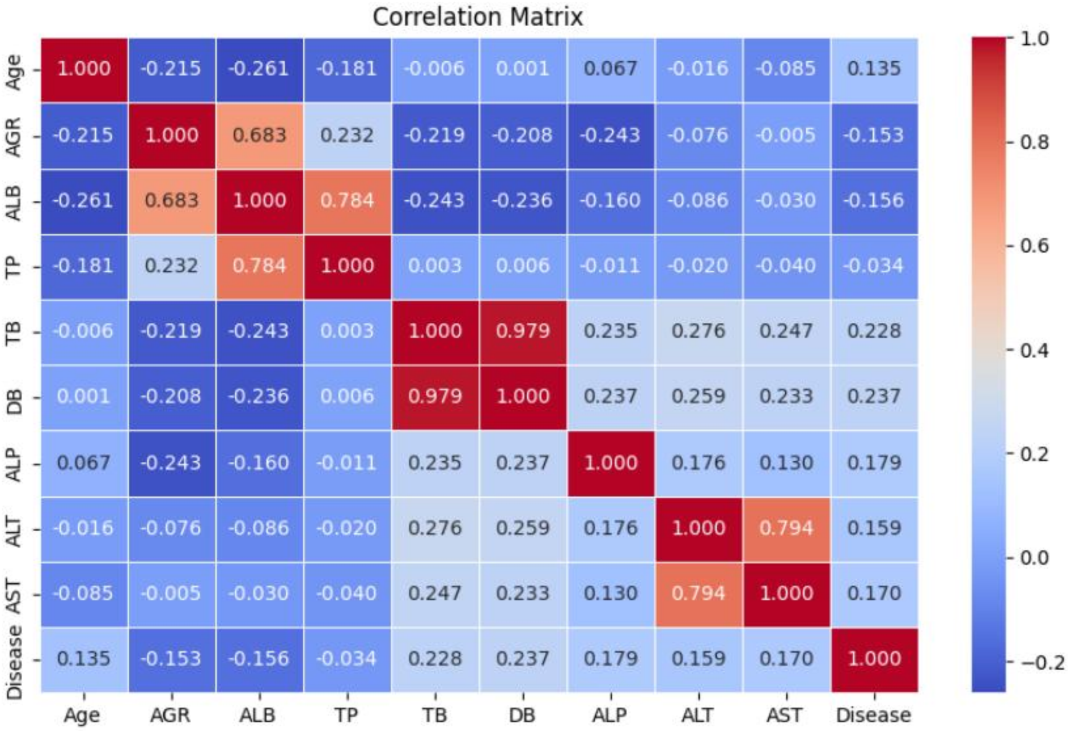


Figure 1.10 Correlation Matrix

Age: Weak negative correlations with AGR (r = -0.215) and ALB (r = -0.261). Positive correlation with Disease (r = 0.135). The correlations suggest a minimal influence of age on the biochemical markers and the presence of liver disease.

Albumin and Globulin Ratio (AGR): Strong positive correlation with ALB (r = 0.683). Weak negative correlation with Disease (r = -0.153). The strong correlation with ALB suggests that as the albumin levels increase, the AGR also increases, indicating a linked relationship between these markers.

Albumin (ALB): Strong positive correlation with TP (r = 0.784) and a moderate positive correlation with AGR (r = 0.683). Weak negative correlation with Disease (r = -0.156). The strong correlation with TP indicates that higher albumin levels are associated with higher total protein levels, which is expected since albumin is a major component of total proteins.

Total Proteins (TP): Strong positive correlation with ALB (r = 0.784). Weak positive correlations with TB (r = 0.003) and DB (r = 0.006). The strong correlation with ALB supports the role of albumin in the total protein measurement. bilirubin, reflecting similar pathological processes affecting both markers.

Total Bilirubin (TB): Strong positive correlation with DB (r = 0.979). Weak positive correlation with Disease (r = 0.228). The strong correlation with DB indicates that direct bilirubin is a significant component of total bilirubin, reflecting similar pathological processes affecting both markers.

Direct Bilirubin (DB): Strong positive correlation with TB (r = 0.979). Weak positive correlation with Disease (r = 0.237). Similar to TB, the strong correlation with TB confirms the intertwined relationship between these two bilirubin markers.

Alkaline Phosphatase (ALP): Weak positive correlations with TB (r = 0.235) and Disease (r = 0.179). The weak correlations suggest ALP levels are somewhat associated with bilirubin levels and the presence of liver disease, but not strongly.

Alanine Aminotransferase (ALT): Strong positive correlation with AST (r = 0.794). Weak positive correlation with Disease (r = 0.159). The strong correlation with AST indicates that these enzymes often rise together in response to liver damage.

Aspartate Aminotransferase (AST): Strong positive correlation with ALT (r = 0.794). Weak positive correlation with Disease (r = 0.170). Similar to ALT, the strong correlation with ALT underscores the parallel changes in these markers in liver pathology.

Disease: Weak positive correlations with TB (r = 0.228), DB (r = 0.237), ALP (r = 0.179), ALT (r = 0.159), and AST (r = 0.170). Weak negative correlations with AGR (r = -0.153) and ALB (r = -0.156). These correlations suggest that while there are associations between these markers and the presence of liver disease, they are relatively weak, indicating the need for a multifactorial approach in diagnosing liver disease.

The correlation matrix reveals significant relationships among various biochemical markers used in the evaluation of liver function. Notably, the strong correlations between ALT and AST, TB and DB, and ALB and TP reflect their interdependent roles in liver physiology and pathology. While some markers show weak correlations with the presence of liver disease, this underscores the complexity of liver disease diagnosis and the necessity of comprehensive diagnostic strategies. The findings provide valuable insights for clinical evaluations and further research into liver disease biomarkers.

* * * * *

The dataset reveals significant variability in the biochemical markers and clinical profiles of the patients. The skewness and kurtosis values indicate that many of the features are not normally distributed, with several features exhibiting high positive skewness and leptokurtosis. This suggests that the majority of the patients have lower values for these markers, with a few patients exhibiting extremely high values.The Shapiro-Wilk test confirms that most features significantly deviate from normality, which may necessitate data transformation or the use of non-parametric statistical methods in subsequent analyses.

The presence of liver disease is relatively high in this dataset, with approximately 69% of the patients diagnosed with liver disease. This imbalance in the target variable should be considered when developing predictive models, as it may impact the performance and evaluation of the models.

Overall, the dataset provides a rich source of information for developing predictive models for liver disease, with a diverse range of biochemical markers and clinical profiles. The significant deviations from normality in many features highlight the need for careful data preprocessing and the potential use of advanced statistical and machine learning techniques to accurately predict liver disease.

2. METHODOLOGY

2.1 Background

This section will mainly outline the steps in preprocessing the data, training the machine learning model and evaluating the performance of those machine learning. All the steps above will be utilizing Python 3.10.12 to process the data, along with a folder which contains object oriented programming (OOP) functions for complicated steps that show in the jupyter notebook interface. OOP allows the code to be easily editable, reusable, shareable among our group members, besides ensuring the process of coding is efficient, reliable and reproducible. (Day, 2024) All the necessary python packages will be installed into the python environment as well.

2.2 Data Preprocessing

Begins by defining a data dictionary, which maps numerical values to categorical labels for various demographic and health-related attributes. Specifically, the dictionary includes mappings for "Ethnic" (0: Malay, 1: Chinese, 2: Indian), "bmi" (0: Normal BMI, 1: Overweight), "Disease" (0: No liver disease, 1: Have Liver Disease), and "Gender" (0: Female, 1: Male). This dictionary will be used later to reassign categorical values in the dataset.

Next, defines a dictionary for renaming certain columns in the dataset to more intuitive names. The columns "Sgot", "Sgpt", and "Alkphos" are renamed to "ALT", "AST", and "ALP", respectively. This renaming is intended to make the column names more consistent with common medical terminology.

Then establishing a dictionary of normal values for various blood test results, including Total Protein (TP), Albumin (ALB), Total Bilirubin (TB), Alkaline Phosphatase (ALP), Alanine Transaminase (ALT), and Aspartate Transaminase (AST). These normal values are provided as ranges and will serve as a reference for interpreting the blood test results in the dataset.

Several variables are then prepared to facilitate the analysis. The dependent variable is identified as "Disease", which indicates the presence or absence of liver disease. The independent variables are divided into demographic and investigation categories. The demographic variables include "Age", "Gender", "Ethnic", and "bmi", while the investigation variables include various blood test results such as "AGR", "ALB", "TP", "TB", "DB", "ALP", "ALT", and "AST". The independent variables are further categorized into continuous variables (which include "Age" and the investigation variables) and categorical variables (which include "Gender", "Ethnic", and "bmi").

The dataset is then loaded from a CSV file located at a specified path. The pd.read_csv function from the pandas library is used to read the CSV file into a DataFrame named df. Once the dataset is loaded, the code proceeds to rename the specified columns using the previously defined renaming dictionary. The rename method of the DataFrame is used to achieve this.

To reassign the categorical values in the dataset, the code iterates over the keys in the data dictionary (excluding "Gender") and adjusts the values in the corresponding columns by subtracting 1. This step ensures that the numerical values in the dataset align with the mappings defined in the data dictionary.

Next step is converting the "Gender" column in the dataset from categorical to numerical values using a label encoding method. The "Gender" column, originally mapped as an object, is now encoded numerically, facilitating easier analysis and modeling.

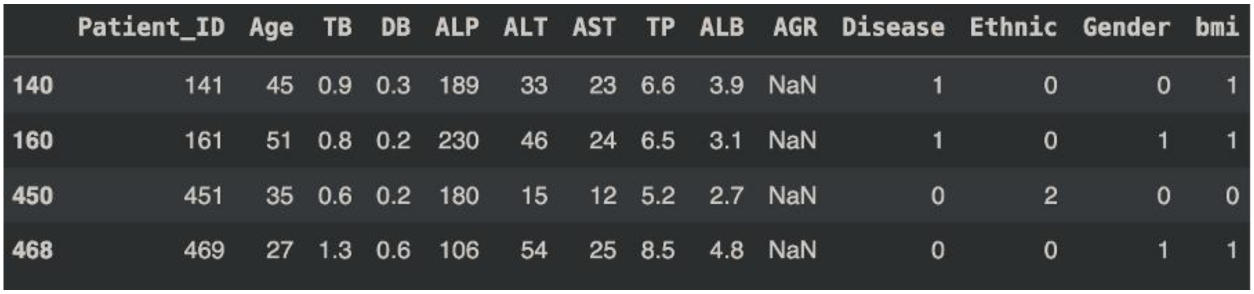
In summary, the code performs a series of preprocessing steps to prepare a dataset for analysis. It defines mappings for categorical values, renames columns for clarity, establishes reference ranges for blood test results, categorizes variables, loads the dataset, adjusts categorical values, and prints a summary of the dataset. These steps are essential for ensuring that the dataset is clean, well-structured, and ready for subsequent analysis.

2.3 Data Cleaning

Next, check for duplicate entries in the dataset. Identifying and handling duplicates is crucial to ensure the integrity and accuracy of the dataset, as duplicate entries can skew analysis results and lead to incorrect conclusions. Following the duplication check, next analyzes the dataset for missing values. The summary includes information on the number of missing values in each row, which is essential for understanding the completeness of the dataset. Handling missing values appropriately is a critical step in data preprocessing, as missing data can impact the performance of machine learning models and the validity of statistical analyses.

The results indicated that no duplicate entries were found in the dataset. This is a positive outcome, as duplicate entries can introduce bias and inaccuracies in the analysis. The absence of duplicates ensures that each row in the dataset represents a unique patient, thereby maintaining the integrity of the data.

Despite during the loading of data no missing data was noticed, but the dtype of AGR showed as object instead of float, by changing ' ' value into null value in the dataframe showed there is 4 missing data in the AGR column, indicating 0.72% of the whole data. The specific data as shown below:

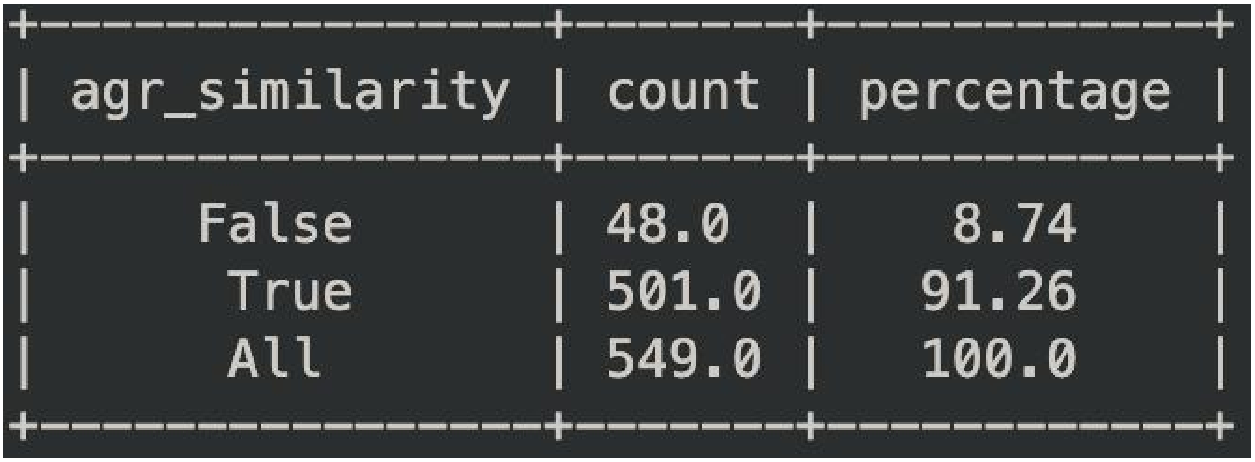


	Patient_ID	Age	TB	DB	ALP	ALT	AST	TP	ALB	AGR	Disease	Ethnic	Gender	bmi
140	141	45	0.9	0.3	189	33	23	6.6	3.9	NaN	1	0	0	1
160	161	51	0.8	0.2	230	46	24	6.5	3.1	NaN	1	0	1	1
450	451	35	0.6	0.2	180	15	12	5.2	2.7	NaN	0	2	0	0
468	469	27	1.3	0.6	106	54	25	8.5	4.8	NaN	0	0	1	1

Figure 2.1 Data with Missing Values for AGR

From the table above, the missing data do not follow a pattern and can be considered as missing completely at random. Due to the nature of how the AGR is calculated by dividing the ALB with Globulin, where Globulin is calculable by subtracting TP with ALB, this will allow us to impute the above missing data with the formula mentioned. However, we are not sure how the data was collected in the first place, therefore it is worthwhile for us to check on the similarity between AGR calculated with this formula with the real data.

The relative tolerance was set to 0.1 due to some of the AGR only having a single decimal while some have 2 decimal, allowing a higher similarity chance between the calculated AGR and the real AGR. However, the results shows around 91.26% of the data were having similar AGR between calculated and real data, the remaining 48 (8.74%) were detected as having different AGR compared to calculated AGR as shown below:



```
+----------------+-------+------------+
| agr_similarity | count | percentage |
+----------------+-------+------------+
|     False      | 48.0  |    8.74    |
|      True      | 501.0 |   91.26    |
|      All       | 549.0 |   100.0    |
+----------------+-------+------------+
```

Figure 2.2 Similarity results between calculated AGR and real AGR

Due to this result above, we proceed with imputing the missing data with the missforest algorithm and Multivariate Imputation by Chained Equations (MICE). Both the imputation results were then compared with the AGR calculated based on the formula above, the imputation result with value near to the calculated AGR were chosen to proceed with the next step.

2.4 Identifying and Handling Outliers

The outliers are identified by calculating the interquartile range (IQR) method. IQR calculated by minus the value between third quantile and first quantile of continuous type of data. Values that fall outside of median add/minus 1.5 times IQR will be considered as outliers. Due to the large amount of the data consisting of outliers, where a total of 187 (33.82%) rows will be affected, therefore it might not be a good choice to remove the outliers. Instead, the outliers will be handled as an imbalance of the class with Synthetic Minority Over-sampling Technique (SMOTE) and feature scaling.

2.5 Calculating Descriptive Statistics

The descriptive statistics will be divided into 2 parts where numerical variables will have mean, standard deviation, minimum, maximum, quartiles calculated. The histogram, QQ plot, box plot were used upon performing descriptive statistics on numerical data. Correlations were calculated to test whether a relationship exists between two variables. Shapiro value was calculated to look for normal distribution of the variable. The categorical data were described as proportions along with visualization of pie chart and chi square test between each of the categorical independent variable and dependent variable were performed to look for their association.

2.6 Dataset Splitting

The dataset were split into 2 different groups, stratified by the outcome variable. Due to the small amount of the dataset, our group decided to use 80% of the data for the training set and the remaining 20% as the test set for evaluation of model performance.

2.7 Feature Scaling and Handling Class Imbalance

Due to the nature of the data where imbalance were noticed, such as more patients without liver disease were noted, the training set were scaled with robust scaler to prevent any single feature from disproportionately influencing the model performance and SMOTE was use to generate synthetic sample for minority class, improve the model's ability to learn from both classes via balanced dataset.

2.8 Model Building

There are total of seven (7) machine learning (ML) algorithm, as shown below:

2.8.1 Logistic Regression

A supervised learning algorithm that uses binary classification for its outcome. The model is able to estimate the coefficients that represent the relationship between the independent variables and the log-odds of the target variable, the probability of the patient having liver disease or not in this model.

2.8.2 Support Vector Machine

A powerful supervised learning algorithm that can be used for both classification and regression tasks by marginalizing the different classes of outcome based on hyperplanes in a high-dimensional space. The decision boundary that separating the patients with or without liver disease can be decided based on the features provided.

2.8.3 k-Nearest Neighbors (kNN)

A non-parametric, instance-based learning algorithm type of machine learning. Based on k nearest neighbors provided in the training data, a new point can be classified by calculating and comparing the average value of those neighbors data to the new data point. This testing set of data will have its position in the feature space and will have its prediction assigned based on its k nearest neighbors value.

2.8.4 Decision Tree

A tree-like model that makes predictions based on a series of binary rules applied to features of the data. This non-parametric supervised learning algorithm starts with a root and develops branches based on decision rules. The branches will end with a node where no further branching is needed or decision rules were able to separate the classes of the outcome, which is either the patient having liver disease or not in this case.

2.8.5 Random Forest

Random Forest is an ensemble learning method that constructs multiple decision trees during training and combines their outputs for prediction. By using bootstrap aggregating (bagging) method, a subset of features for each tree being introduced randomly to reduce overfitting of the model. Random Forests are robust to noise and can handle high-dimensional data.

2.8.6 Light Gradient Boosting Machine (LightGBM)

A highly efficient and scalable gradient boosting framework which utilizes Gradient-based One Side Sampling (GOSS) method to remove the information gain with data instances that have little impact during the tree construction and use leaf-wise method in making the prediction. This provides a faster training speed, low memory usage yet high accuracy in predicting the probability of a patient having liver disease or not in this scenario.

2.8.7 Extreme Gradient Boosting (XGBoost)

Another highly efficient and scalable implementation model of the gradient boosting framework. XGBoost builds the ensemble of weak decision tree models in a sequential manner, where each new tree aims to correct the errors of the previous trees. The model is also capable to improving performance via parallel processing, tree pruning and regularization as well.

2.9 Model Evaluation

All the model above will be evaluated based on time required to construct the model, feature importances, cross validation, “Area Under the Curve” of the “Receiver Operating Characteristic” curve (AUC-ROC), performance metrics across different thresholds, confusion matrix, classification report and some metrics, such as accuracy, precision, sensitivity, specificity, F1 score, mean squared error (MSE), mean absolute relative difference (MARD), recall score, cross-validation score (CV Score) etc. This will provide us to understand the strength and weakness of the model in predicting the presence of liver disease of the patient or not.

For example, the training time indicating the cost of computation and its scalability, it is important while dealing with time-sensitive tasks such as payment to providers based on the prediction and diagnosis provided by the provider. The feature selection allows us to select the best feature in forming the model and reduce noise, improve model interpretability as well. Cross validation allows us to examine the generalizability of the model. AUC-ROC being able to distinguish between positive and negative cases, along with performance at different thresholds will improve the model performance by controlling the trade-off between true positive and false positive values of the model.

Various performance of the model can also be visualized via confusion matrix and classification report. The accuracy of the model measures the proportion of correct predictions made by the model. Precision measures only the true positives among all positive prediction. Sensitivity measure the percentage of actual positive cases the model correctly identified while specify measure how well the model identified the true negative cases. Climbing the precision and recall will form a single score call as F1 score. MSE quantifies the overall prediction error by measuring the average squared difference between predicted and actual values. Meanwhile the MARD provides an interpretable error measure by calculating the average absolute relative difference between predicted and actual values.

All the above measures being used in choosing the best model of machine learning. Due to its unseen future role, the model chosen based on two (2) different scenarios, which is for screening of disease and for payment of providers. The first scenario will be useful to alert the provider a patient might have liver disease and need further investigation, which will require high sensitivity. For the second scenario, where the provider submitted a medical claim will need to be screened on whether the information is true or not, thus requiring a high specificity and fast training model to deal with.

3. RESULT

The performance of the seven (7) models was evaluated on the test set.

3.1 Logistic Regression

The analysis presented in Figure 3.1 suggests that the model performs well in distinguishing between classes, as evidenced by the AUC of 0.8209. The optimal threshold for the model is identified at 0.4167 (Figure 3.2), where it achieves the highest accuracy while maintaining a good balance between sensitivity and specificity. The feature importance chart highlights that 'bmi', 'Age', and 'DB' are critical (Figure 3.3) in the model's decision-making process, suggesting that these factors should be closely monitored or further investigated in the context of the model's application.

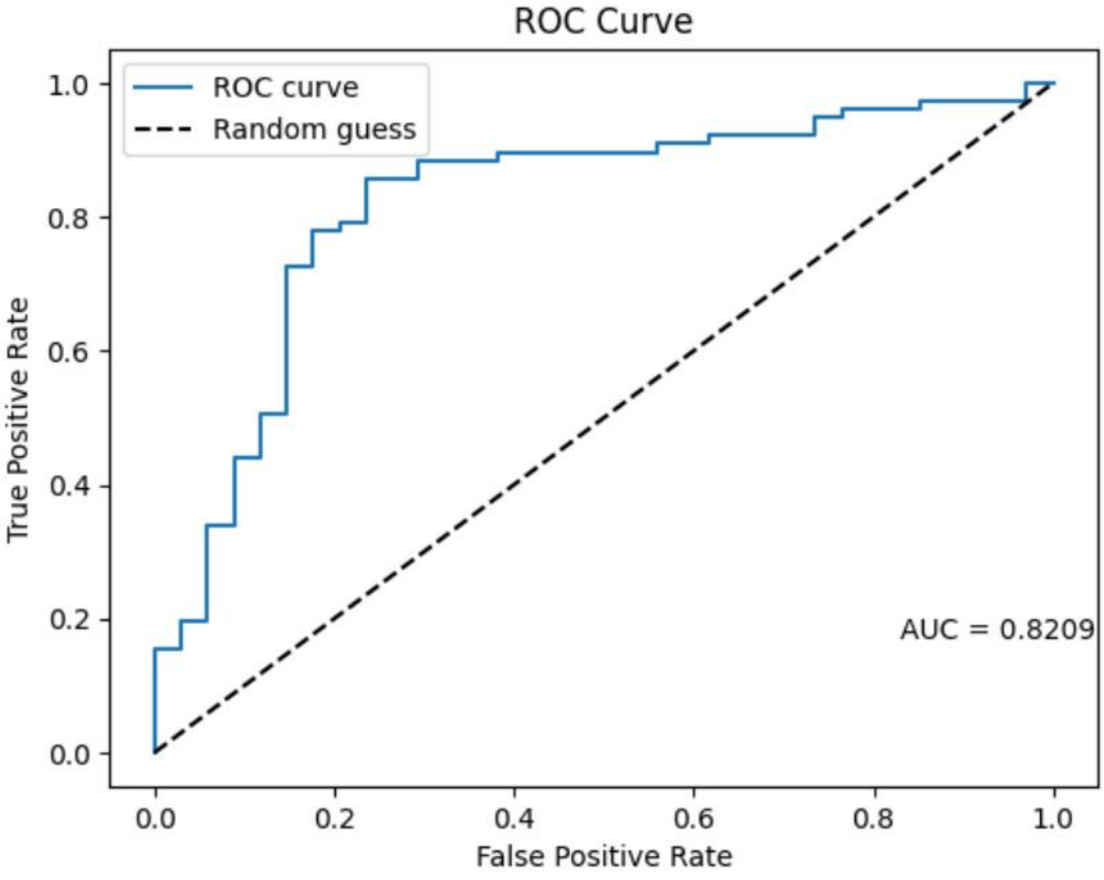


Figure 3.1 ROC Curve for Logistic Regression

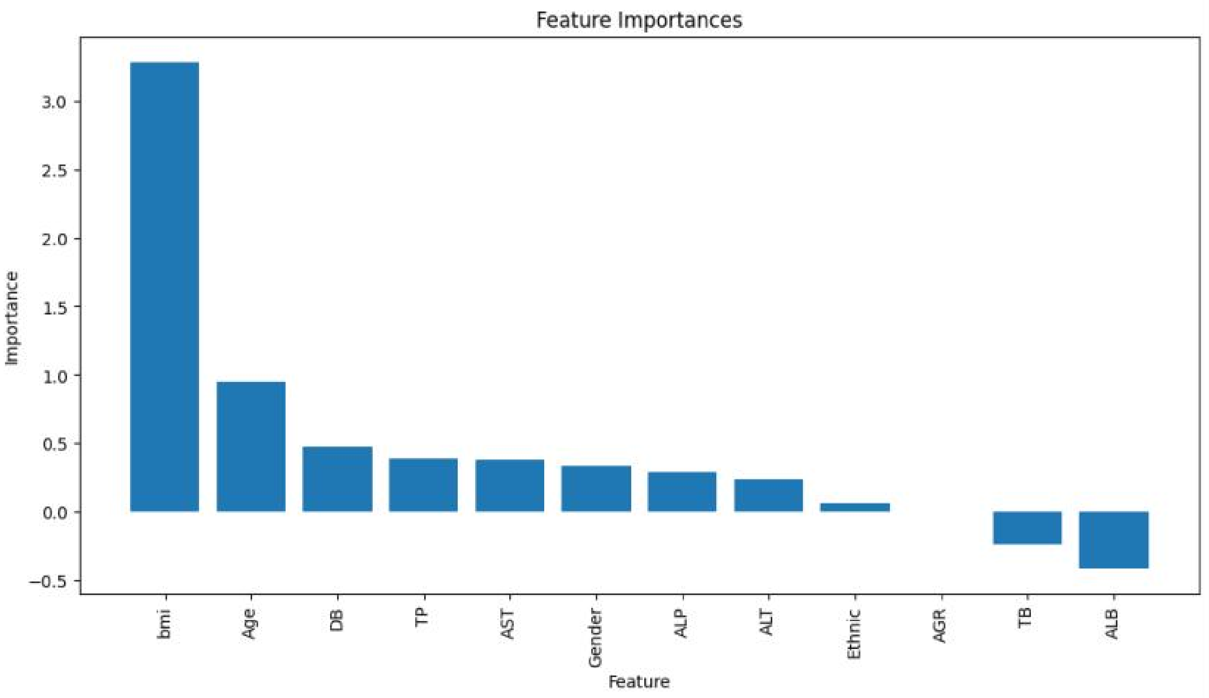


Figure 3.2 Feature Importance in Predictive Model for Logistic Regression

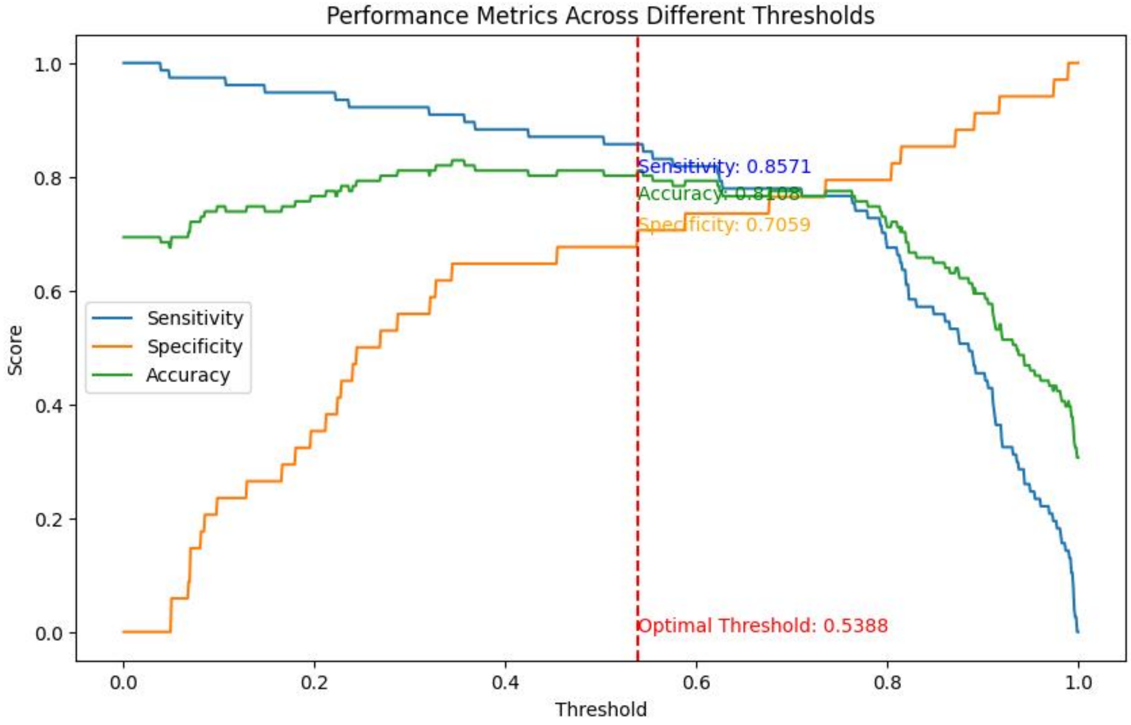


Figure 3.3 Performance Metrics Across Different Thresholds for Logistic Regression

The evaluation of the logistic regression model, as evidenced by the confusion matrix (Table 3.1) and classification report (Table 3.2), indicates a robust performance in classifying conditions as either true or false. The confusion matrix reveals that the model correctly predicted the true condition 66 times out of 77, and the false condition 8 times out of 34, resulting in an overall accuracy of 83%. This level of accuracy is commendable and suggests that the model is effective in distinguishing between the two conditions.

Table 3.1 Confusion Matrix for Logistic Regression

Condition	Correct Prediction	Wrong Prediction	All
False Condition	8	26	34
True Condition	66	11	77
All	74	37	111

The classification report further supports these findings, with a precision of 0.89 for the true condition and 0.70 for the false condition, indicating a high likelihood that the model's predictions are correct when it predicts a true condition. The recall scores of 0.86 for the true condition and 0.76 for the false condition suggest that the model is reasonably capable of identifying all relevant instances of each condition. The F1-scores, which balance precision and recall, are 0.87 for the true condition and 0.73 for the false condition, further indicating good model performance.

Moreover, the model's performance metrics such as the AUC of 0.8209 and the optimal threshold of 0.4167, which maximizes sensitivity and specificity, highlight its efficiency in operational settings. The time required for model training and prediction was approximately 4.83 seconds, demonstrating not only the model's effectiveness but also its efficiency.

Table 3.2 Classification Report for Logistic Regression

Class	Precision	Recall	F1-Score	Support
0	0.70	0.76	0.73	34
1	0.89	0.86	0.87	77
Accuracy			0.83	111
Macro Avg	0.80	0.81	0.80	111
Weighted Avg	0.83	0.83	0.83	111

The logistic regression model exhibits a high degree of accuracy, precision, and recall, making it a reliable tool for binary classification tasks. The balance between sensitivity and specificity, as optimized by the model's threshold setting, provides a practical approach for real-world applications where decision-making processes depend on accurate and timely predictions.

3.2 Support Vector Machine (SVM)

The AUC value of 0.8514, as shown in Figure 3.4, demonstrates excellent predictive capability. The feature 'bmi' stands out as the most influential, as depicted in Figure 3.5, suggesting it plays a critical role in the model's predictions and should be prioritized in future analyses or model enhancements. Figure 3.6 illustrates the model's sensitivity, specificity, and accuracy across various thresholds, with the optimal threshold identified by a red dashed line at approximately 0.6568. At this point, the model achieves a balance between sensitivity and specificity, with sensitivity approximately at 0.8312 and specificity around 0.8529, as highlighted by the intersections with the dashed line.

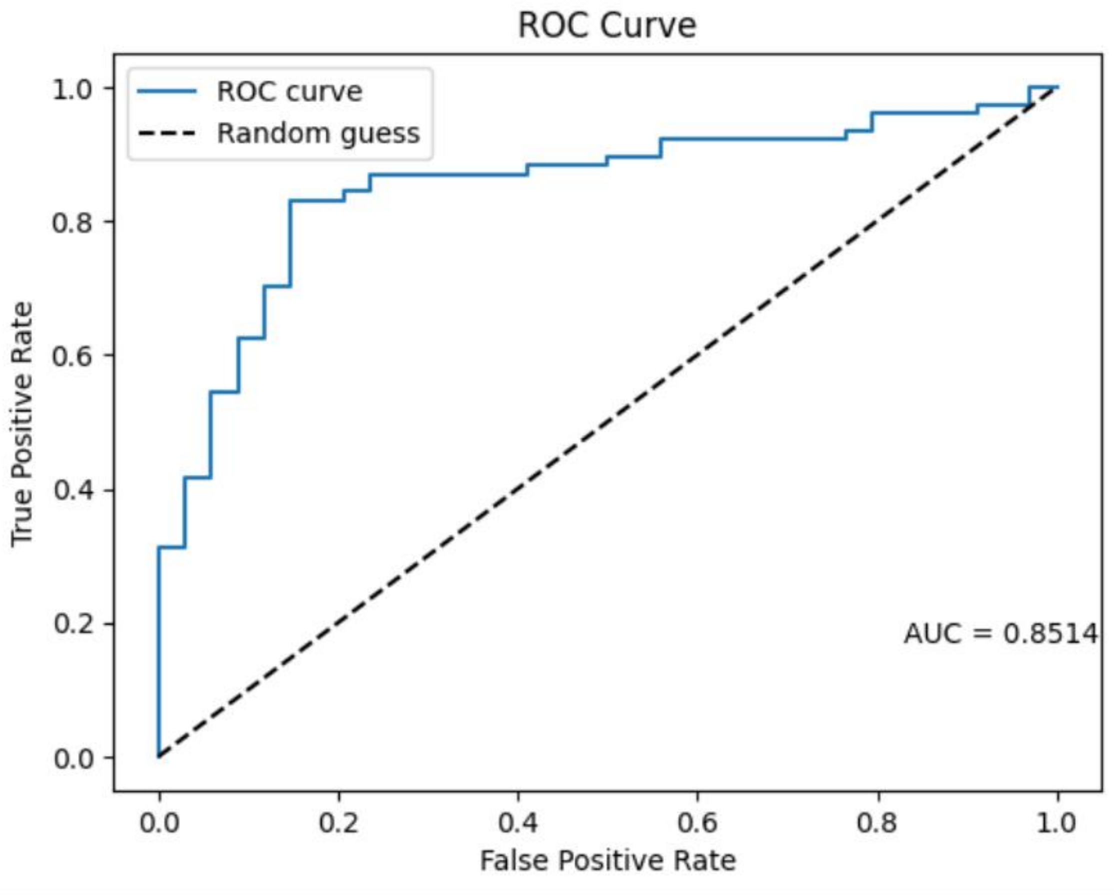


Figure 3.4 ROC Curve for Support Vector Machine

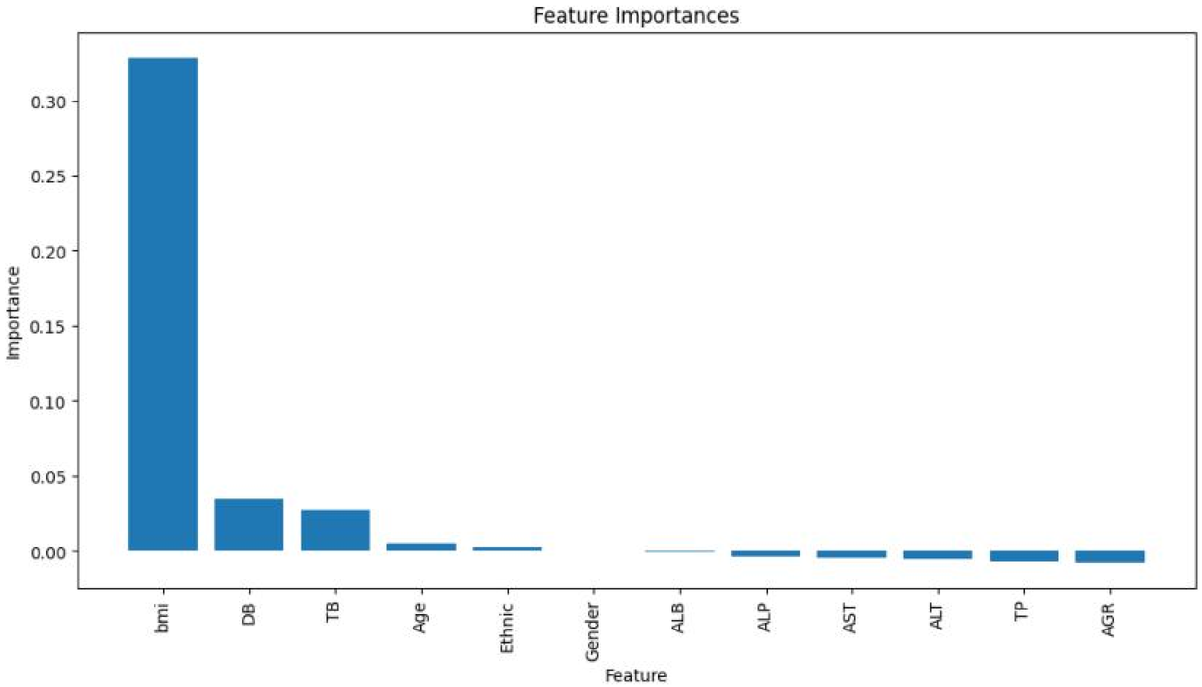


Figure 3.5 Feature Importance in Predictive Model for Support Vector Machine

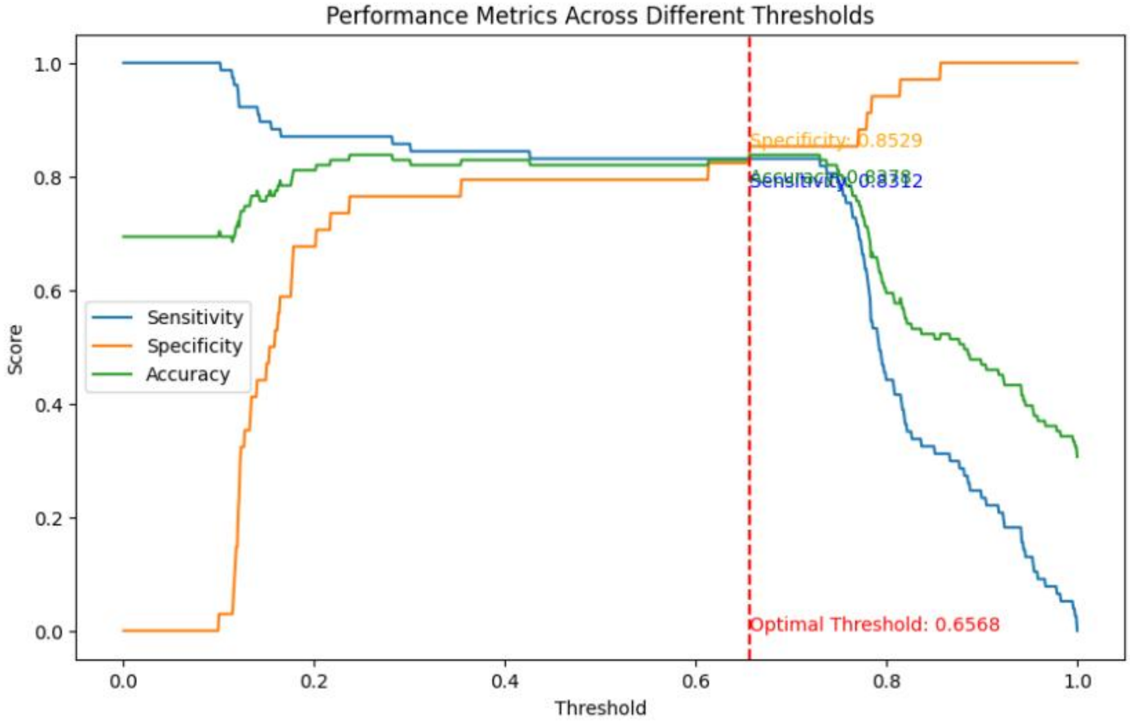


Figure 3.6 Performance Metrics Across Different Thresholds for Support Vector Machine

The confusion matrix for the SVM model (Table 3.3) provides a detailed breakdown of the model's predictive accuracy across different conditions. The matrix reveals that the model correctly identified the true condition 64 times (True Positives) and correctly predicted the false condition 5 times (True Negatives). However, the model also made some errors, incorrectly predicting the true condition 29 times when it was actually false (False Positives) and failing to identify the true condition 13 times when it was indeed true (False Negatives). This resulted in a total of 34 actual false conditions and 77 actual true conditions among the 111 observations.

Table 3.3 Confusion Matrix for Support Vector Machine

Condition	Correct Prediction	Wrong Prediction	All
False Condition	5	29	34
True Condition	64	13	77
All	69	42	111

The SVM model demonstrates strong predictive performance (Table 3.4), particularly in identifying true conditions (class 1) with high precision. The high F1-score for class 1 suggests that the model is well-tuned for this class, balancing the trade-off between precision and recall effectively. However, the model struggles somewhat with false conditions (class 0), as evidenced by a lower precision and a higher rate of false positives.

The overall accuracy of 0.84 indicates that the model is robust, but there may be room for improvement, especially in reducing false positives and increasing true negatives. This could potentially be addressed by further tuning the model's parameters, gathering more balanced data, or applying different feature selection techniques to enhance the model's ability to generalize across different conditions.

Table 3.4 Classification Report for Support Vector Machine

Class	Precision	Recall	F1-Score	Support
0	0.69	0.85	0.76	34
1	0.93	0.83	0.88	77
Accuracy			0.84	111
Macro Avg	0.81	0.84	0.82	111
Weighted Avg	0.85	0.84	0.84	111

In conclusion, the SVM model is effective but could benefit from adjustments to better handle false condition predictions. This analysis should be considered in the context of the specific application and the costs associated with different types of prediction errors. Further studies could explore the impact of these adjustments on the model's performance in real-world scenarios.

3.3 K-Nearest Neighbour(kNN)

The ROC curve presented in Figure 3.7, with an AUC of 0.8654, demonstrates that the model has good predictive accuracy. Figure 3.8 highlights the significance of various features in the model. The most influential feature is 'bmi', indicating that Body Mass Index is a critical predictor in the model. This suggests that 'bmi' holds substantial weight in the decision-making process of the classifier. Other features like 'Ethnic' and 'ALP' also contribute to the predictions but to a lesser extent. The optimal threshold identified is around 0.4897, where the model achieves a balance between sensitivity and specificity, optimizing overall accuracy (Figure 3.9).

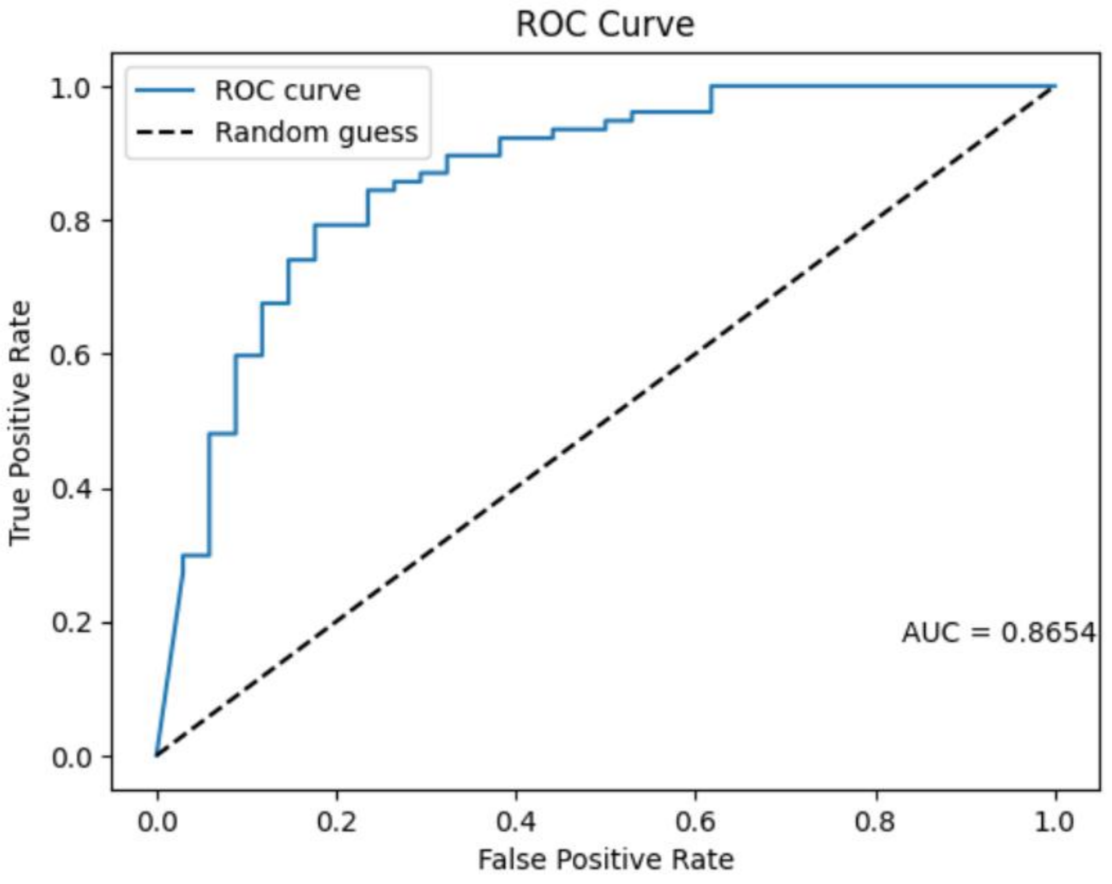


Figure 3.7 ROC Curve for K-Nearest Neighbour

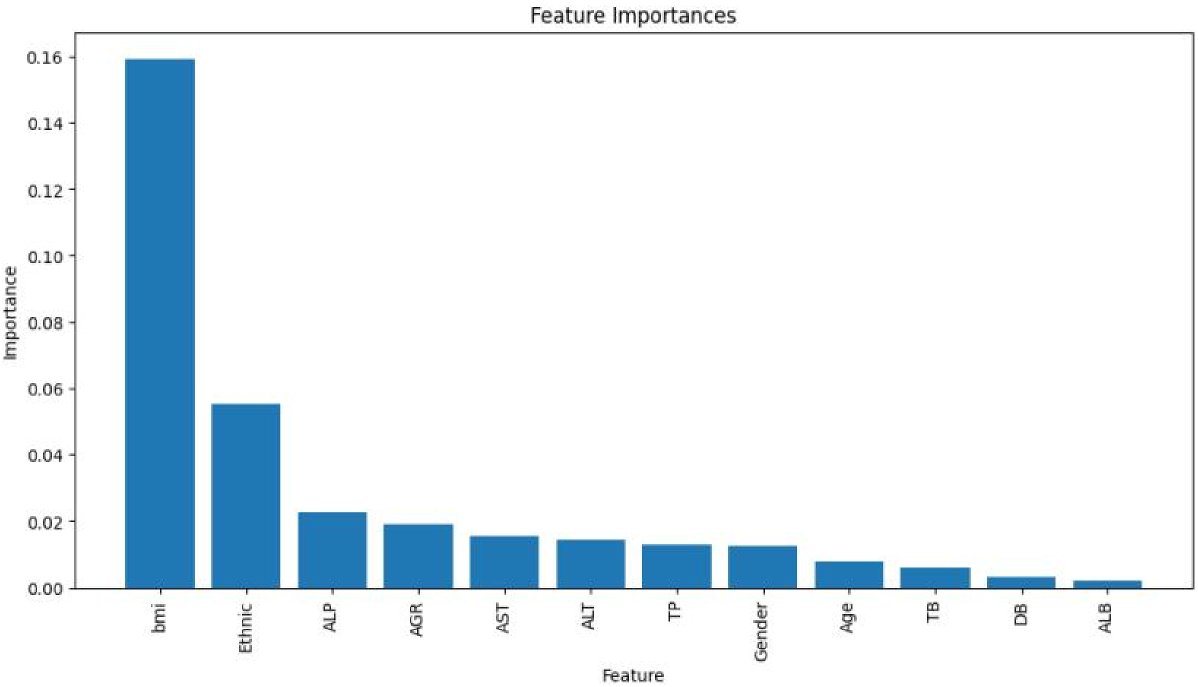


Figure 3.8 Feature Importance in Predictive Model for K-Nearest Neighbour

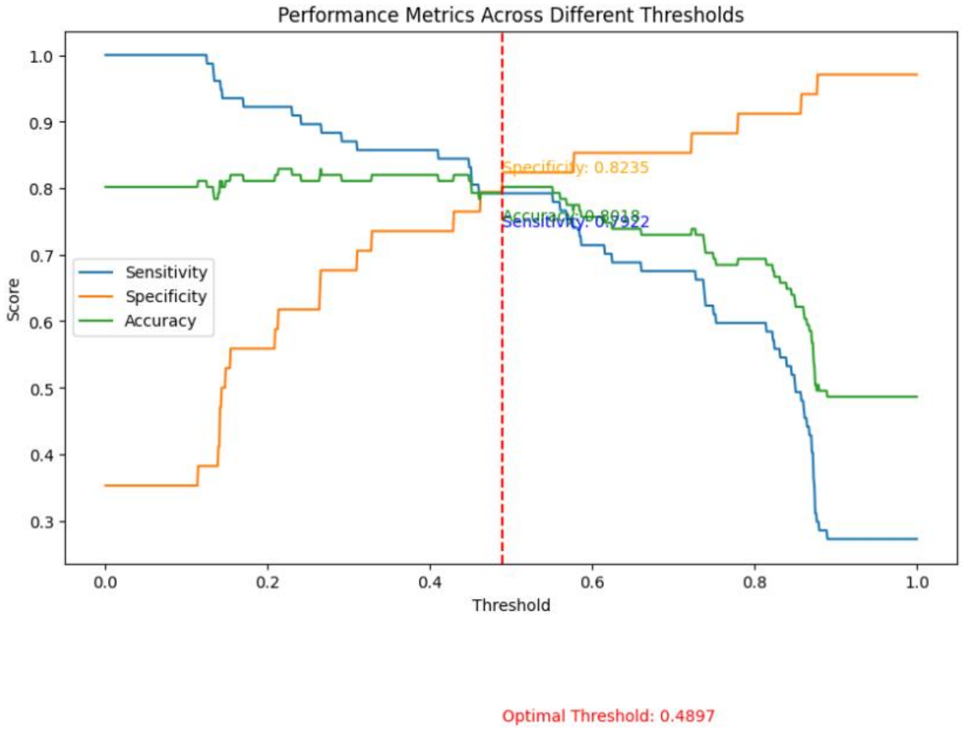


Figure 3.9 Performance Metrics Across Different Thresholds for K-Nearest Neighbour

The confusion matrix provides a breakdown of the classifier's predictions compared to the actual conditions (Table 3.5). For the false condition (class 0), the model correctly predicted 6 instances but incorrectly predicted 28 instances, leading to a total of 34 instances for this class. For the true condition (class 1), the model correctly predicted 61 instances while incorrectly predicting 16, with a total of 77 instances for this class. The overall accuracy can be calculated by the sum of true positives and true negatives divided by the total number of cases, which in this case is about 60.36%.

Table 3.5 Confusion Matrix for K-Nearest Neighbour

Condition	Correct Prediction	Wrong Prediction	All
False Condition	6	28	34
True Condition	61	16	77
All	67	44	111

The model shows a higher precision in predicting the true condition (class 1) compared to the false condition (class 0). This suggests that the model is more reliable when predicting the presence of the condition. The recall for class 0 is higher than for class 1, indicating that the model is better at identifying all relevant instances of class 0.

The overall accuracy of 80% is quite robust, showing that the model performs well across both classes. However, the accuracy calculated from the confusion matrix (about 60.36%) suggests there might be a discrepancy that needs further investigation, possibly due to the interpretation of the matrix or additional factors not accounted for in the accuracy score in Table 3.6.

The F1-scores indicate a good balance between precision and recall, especially for class 1, which has a higher F1-score, suggesting effective classification despite the challenges of balancing type I and type II errors.

Table 3.6 Classification Report for K-Nearest Neighbour

Class	Precision	Recall	F1-Score	Support
0	0.64	0.82	0.72	34
1	0.91	0.79	0.85	77
Accuracy			0.80	111
Macro Avg	0.77	0.81	0.78	111
Weighted Avg	0.83	0.80	0.81	111

In conclusion, the kNN classifier demonstrates a commendable ability to distinguish between the two conditions, with particular strength in predicting the true condition accurately. These findings could be instrumental in refining the model further, especially by addressing the lower recall in class 1 and exploring the reasons behind the discrepancy in the calculated accuracies. This analysis not only aids in understanding the model's current capabilities but also highlights areas for potential improvement in future iterations.

3.4 Decision Tree

In Figure 3.10, AUC is reported as 0.7844, indicating that the classifier possesses a good level of discrimination ability, significantly better than random guessing. This suggests that the model is effective in distinguishing between positive and negative classes. The most important feature is 'bmi', which has a significantly higher importance score compared to other features (Figure 3.11). Other important features include 'DB', 'Age', 'AST', 'TB', 'ALP', 'ALT', 'Gender', 'Ethnic', 'ALB', and 'TP'. Figure 3.12 identifies an optimal threshold at approximately 0.2676, where the sensitivity is about 0.8571, accuracy is around 0.8198, and specificity is approximately 0.7353.

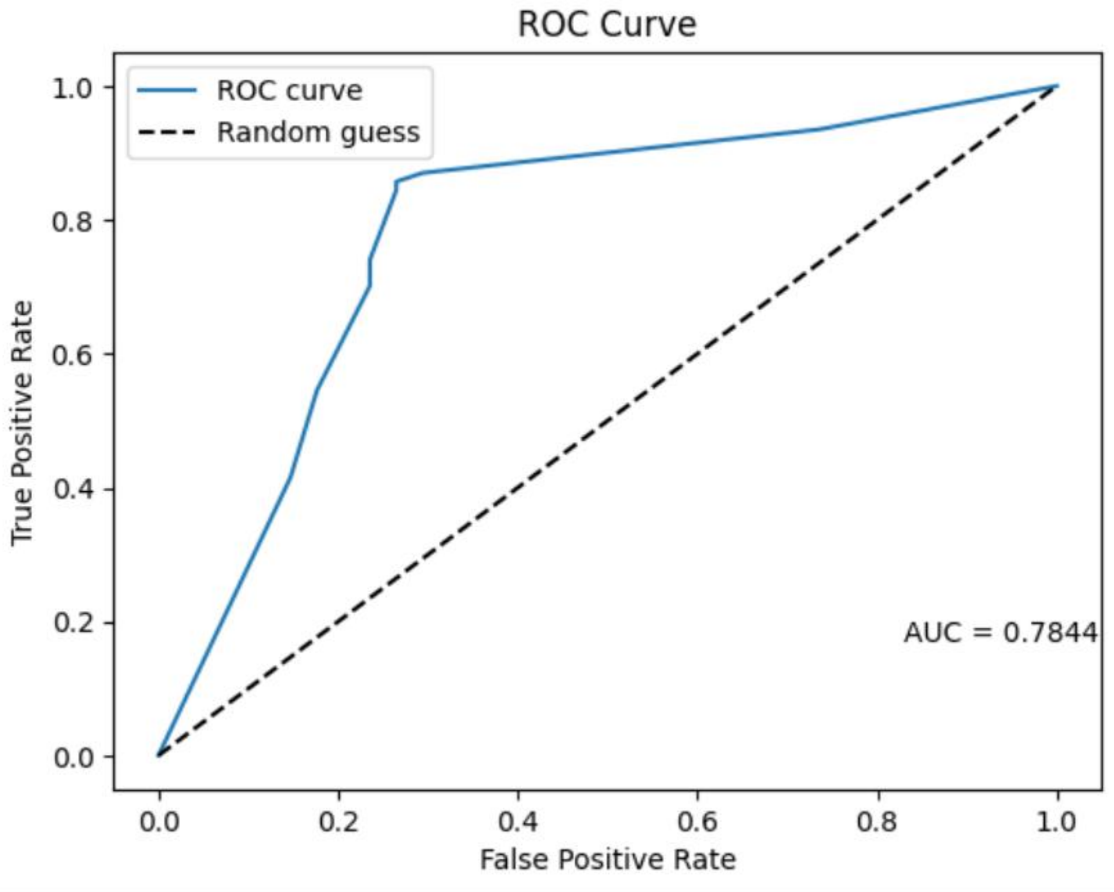


Figure 3.10 ROC Curve for Decision Tree

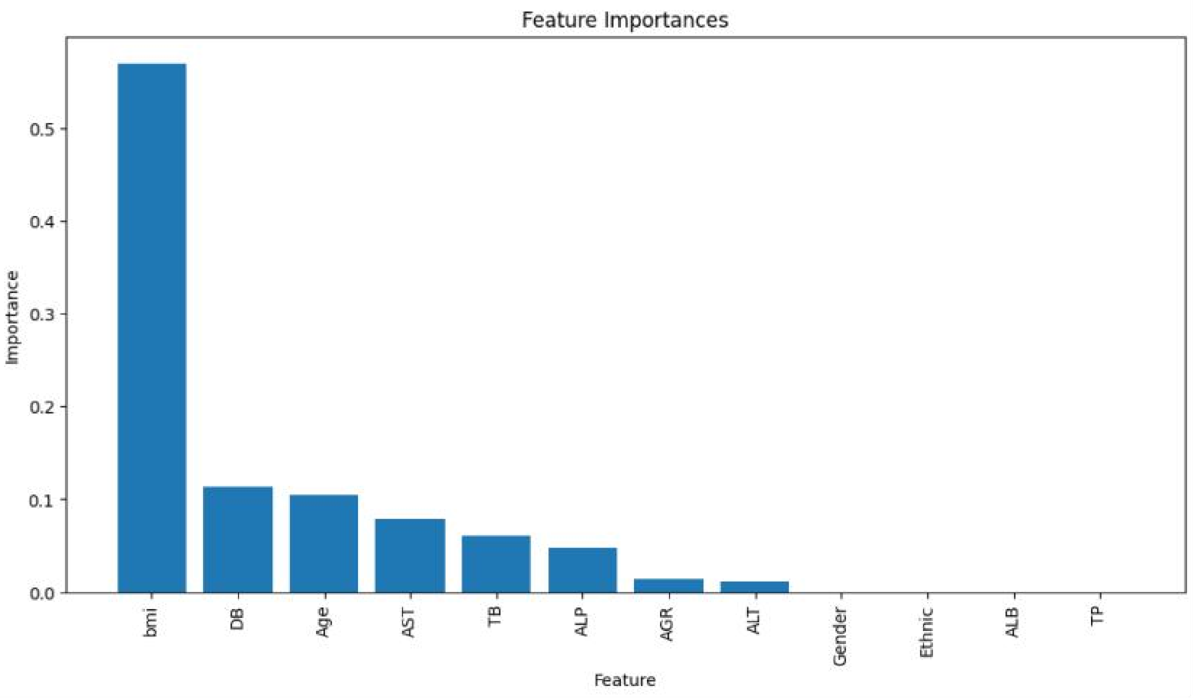


Figure 3.11 Feature Importance in Predictive Model for Decision Tree

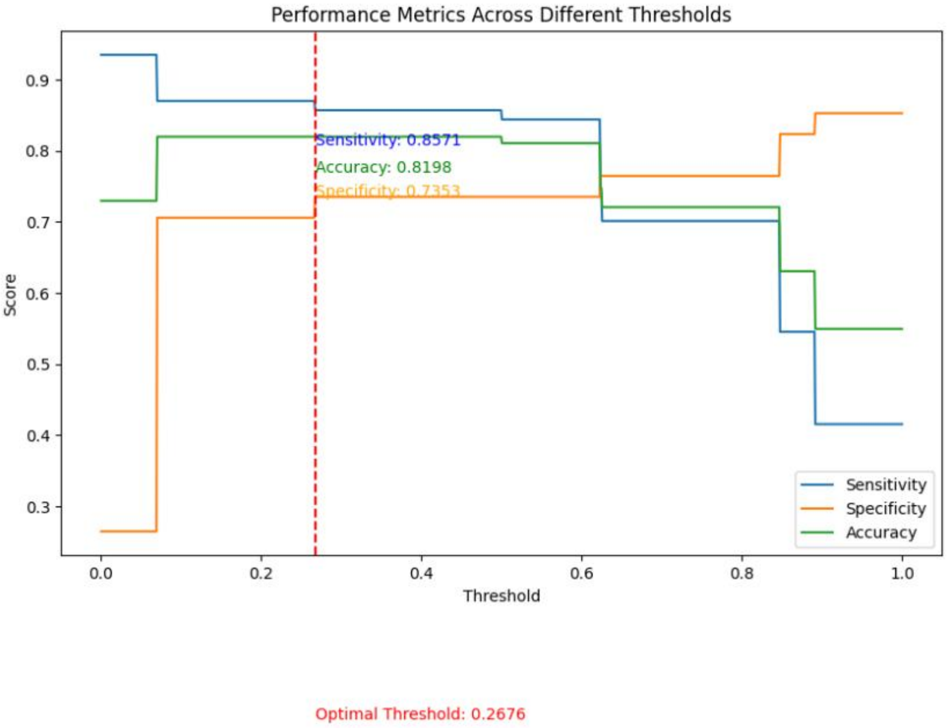


Figure 3.12 Performance Metrics Across Different Thresholds for Decision Tree

The confusion matrix (Table 3.7) and classification report (Table 3.8) provide a comprehensive evaluation of the decision tree model's performance. The confusion matrix reveals that the model has a higher number of false positives (25) compared to false negatives (11), indicating that the model is more likely to incorrectly classify negative instances as positive. This is reflected in the precision and recall values for each class.

For the negative class (Class 0), the precision is 0.69, meaning that 69% of the instances predicted as negative are actually negative. The recall is 0.74, indicating that 74% of the actual negative instances are correctly identified by the model. The F1-score, which is the harmonic mean of precision and recall, is 0.71, suggesting a balanced performance for the negative class.

For the positive class (Class 1), the precision is 0.88, meaning that 88% of the instances predicted as positive are actually positive. The recall is 0.86, indicating that 86% of the actual positive instances are correctly identified by the model. The F1-score for the positive class is 0.87, demonstrating a strong performance in identifying positive instances.

Table 3.7 Confusion Matrix for Decision Tree

Condition	Correct Prediction	Wrong Prediction	All
False Condition	9	25	34
True Condition	66	11	77
All	75	36	111

The overall accuracy of the model is 0.82, indicating that 82% of the total instances are correctly classified. The macro average and weighted average metrics provide additional insights into the model's performance across both classes. The macro average precision, recall, and F1-score are 0.79, 0.80, and 0.79, respectively, indicating a balanced performance across both classes. The weighted average precision, recall, and F1-score are all 0.82, reflecting the overall performance of the model, taking into account the support (number of instances) for each class.

Table 3.8 Classification Report for Decision Tree

Class	Precision	Recall	F1-Score	Support
0	0.69	0.74	0.71	34
1	0.88	0.86	0.87	77
Accuracy			0.82	111
Macro Avg	0.79	0.80	0.79	111
Weighted Avg	0.82	0.82	0.82	111

In conclusion, the decision tree model demonstrates a strong performance, particularly in identifying positive instances, with an overall accuracy of 82%. The model's precision, recall, and F1-score values indicate a balanced performance across both classes, with a slightly higher tendency to incorrectly classify negative instances as positive. These findings suggest that the model is effective in its predictive capabilities, but further refinement may be needed to reduce the number of false positives.

3.5 Random Forest

The ROC curve and AUC value (Figure 3.13) indicate that the model is effective in distinguishing between the classes. With an AUC of 0.8400, the model demonstrates good predictive performance, significantly surpassing the performance of a random guess (AUC = 0.5). Figure 3.14 illustrates the importance of various features in the model, with the 'bmi' feature being the most influential, significantly impacting the model's predictions. Figure 3.15 identifies an optimal threshold at 0.6358, where the balance between sensitivity (0.8571) and accuracy (0.8571) is optimized, as indicated by the red dashed line.

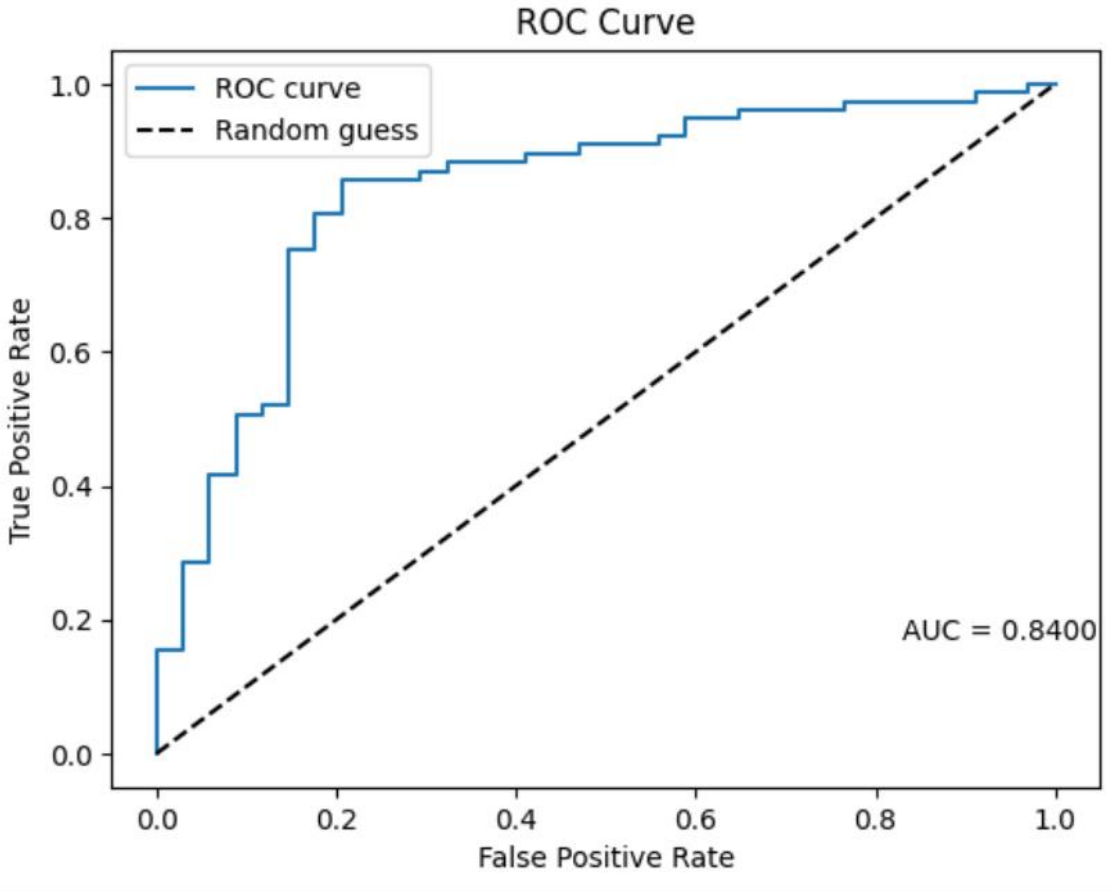


Figure 3.13 ROC Curve for Random Forest

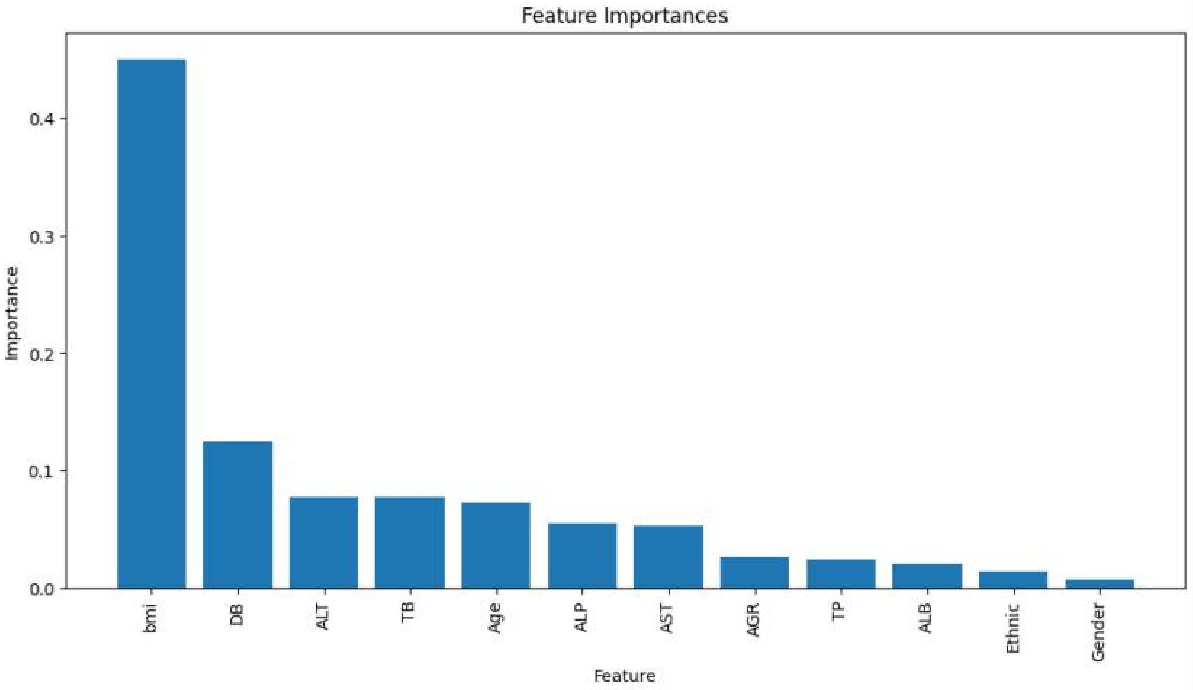


Figure 3.14 Feature Importance in Predictive Model for Random Forest

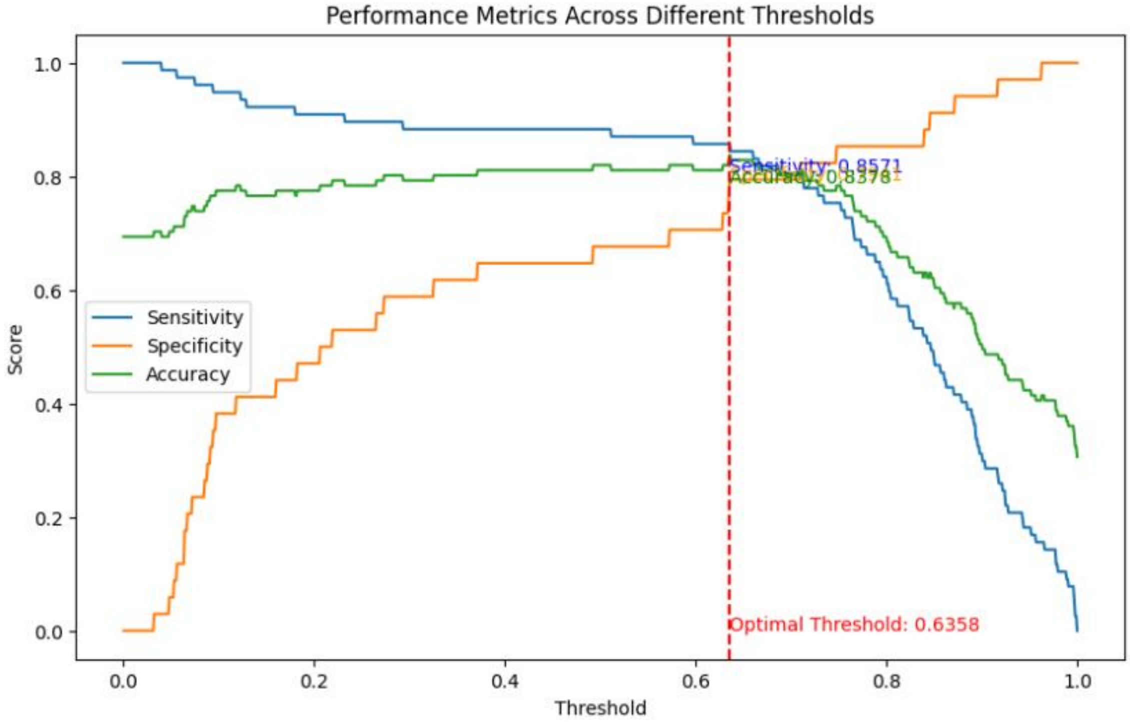


Figure 3.15 Performance Metrics Across Different Thresholds for Random Forest

The confusion matrix (Table 3.9) presents the counts of correct and incorrect predictions made by the Random Forest model. For the "False Condition" (negative class), the model correctly predicted 7 instances and incorrectly predicted 27 instances, out of a total of 34 instances. For the "True Condition" (positive class), the model correctly predicted 66 instances and incorrectly predicted 11 instances, out of a total of 77 instances. Overall, the model made 73 correct predictions and 38 incorrect predictions out of 111 total instances.

Table 3.9 Confusion Matrix for Random Forest

Condition	Correct Prediction	Wrong Prediction	All
False Condition	7	27	34
True Condition	66	11	77
All	73	38	111

The classification report (Table 3.10) provides a more granular view of the model's performance metrics, including precision, recall, F1-score, and support for each class. The high precision (0.90) and recall (0.86) for the positive class suggest that the model is particularly adept at identifying true positive instances, which is crucial in many real-world applications where the cost of false negatives can be high. The balanced F1-scores for both classes indicate that the model maintains a good trade-off between precision and recall, ensuring reliable performance across different metrics.

The overall accuracy of the model is 0.84, indicating that 84% of the total predictions were correct. The macro average, which is the unweighted mean of the precision, recall, and F1-score for both classes, shows values of 0.81, 0.83, and 0.81, respectively. The weighted average, which takes into account the support (number of true instances) for each class, shows values of 0.84 for precision, recall, and F1-score, indicating a consistent performance across both classes.

Table 3.10 Classification Report for Random Forest

Class	Precision	Recall	F1-Score	Support
0	0.71	0.79	0.75	34
1	0.90	0.86	0.88	77
Accuracy			0.84	111
Macro Avg	0.81	0.83	0.81	111
Weighted Avg	0.84	0.84	0.84	111

In conclusion, the Random Forest model exhibits strong predictive capabilities, particularly for the positive class, making it a valuable tool for applications requiring accurate classification.

3.6 Light Gradient Boost Machine (LightGBM)

The ROC curve in Figure 3.16 demonstrates the model's effectiveness in distinguishing between classes, with an AUC of 0.8220, indicating good predictive performance compared to random guessing. Figure 3.17 lists the features included in the model. Age is identified as the most significant predictor, followed by 'DB' and 'TP', highlighting their strong influence on the model's outcomes. Figure 3.18 identifies an optimal threshold at approximately 0.5388, where sensitivity is around 0.8571, accuracy peaks at 0.8570, and specificity is about 0.7059.

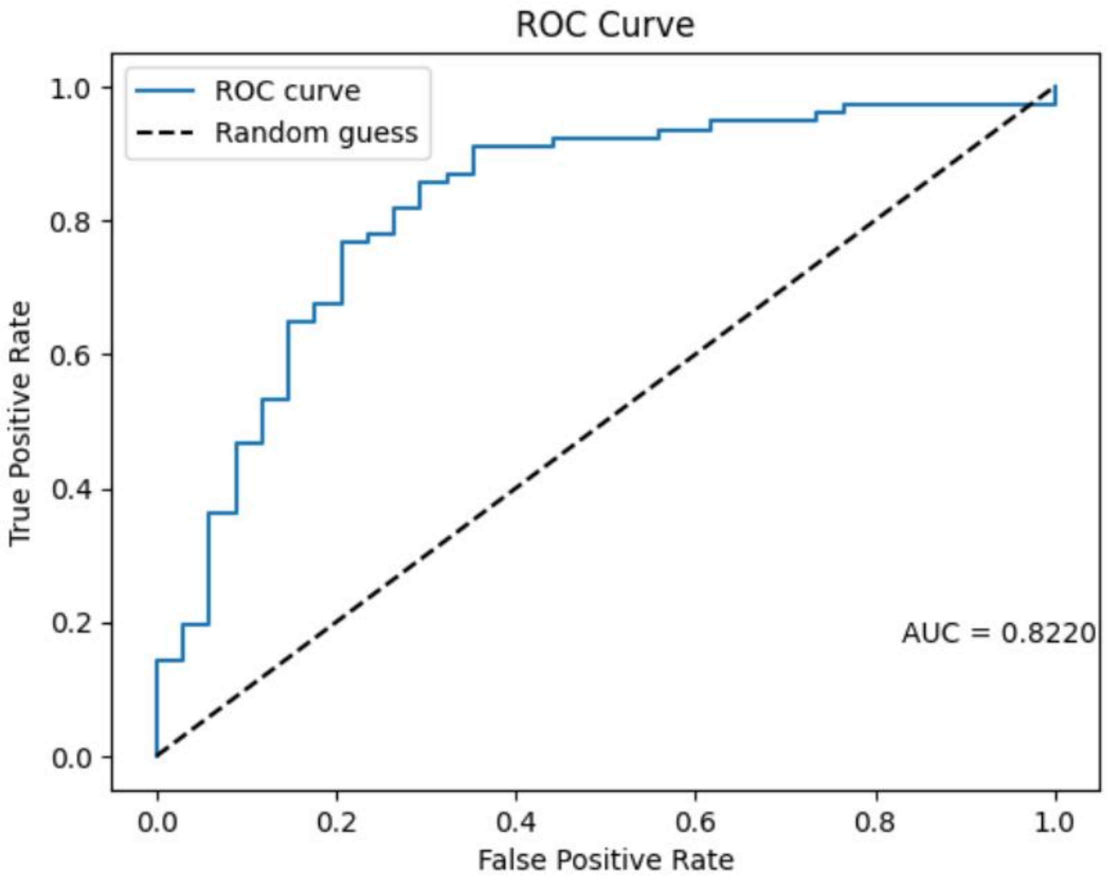


Figure 3.16 ROC Curve for Light Gradient Boost Machine

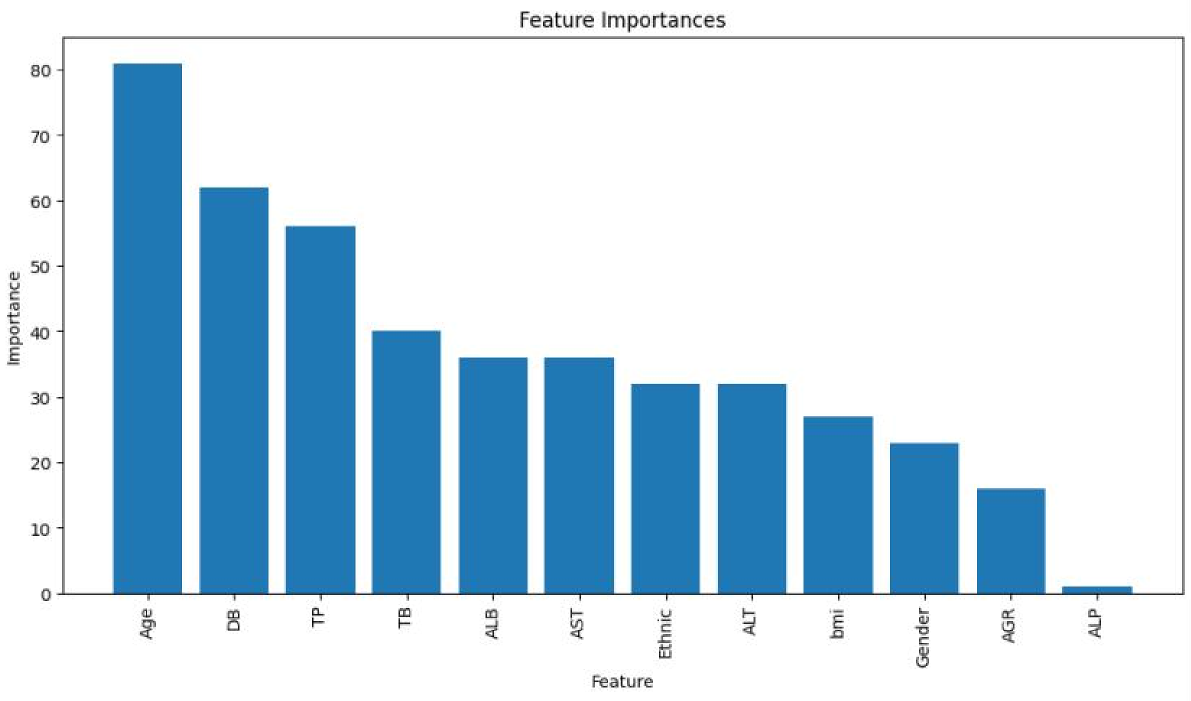


Figure 3.17 Feature Importance in Predictive Model for Light Gradient Boost Machine

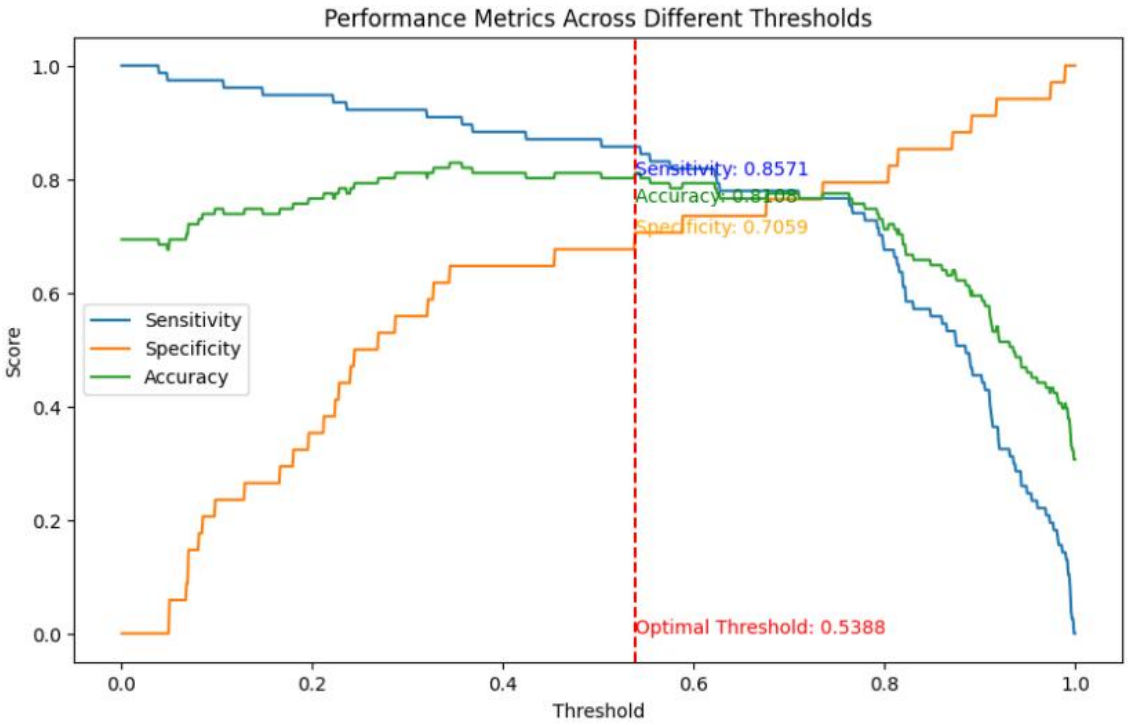


Figure 3.18 Performance Metrics Across Different Thresholds for Light Gradient Boost Machine

Table 3.11 provides the confusion matrix for the LightGBM model. The matrix reveals that out of 111 total instances, the model correctly predicted 76 instances and incorrectly predicted 35 instances. Specifically, for the false condition, the model correctly identified 10 instances and incorrectly identified 24 instances. For the true condition, the model correctly identified 66 instances and incorrectly identified 11 instances. This indicates that the model has a higher accuracy in predicting true conditions compared to false conditions.

Table 3.11 Confusion Matrix for Light Gradient Boost Machine

Condition	Correct Prediction	Wrong Prediction	All
False Condition	10	24	34
True Condition	66	11	77
All	76	35	111

Table 3.12 presents the classification report for the LightGBM model, detailing the precision, recall, F1-score, and support for each class. For class 0 (false condition), the precision is 0.69, recall is 0.71, and F1-score is 0.70, with a support of 34 instances. For class 1 (true condition), the precision is 0.87, recall is 0.86, and F1-score is 0.86, with a support of 77 instances. The overall accuracy of the model is 0.81, indicating that the model correctly predicts 81% of the instances. The macro average for precision, recall, and F1-score is 0.78, while the weighted average for these metrics is 0.81.

Table 3.12 Classification Report for Light Gradient Boost Machine

Class	Precision	Recall	F1-Score	Support
0	0.69	0.71	0.70	34
1	0.87	0.86	0.86	77
Accuracy			0.81	111
Macro Avg	0.78	0.78	0.78	111
Weighted Avg	0.81	0.81	0.81	111

The findings indicate that the LightGBM model performs well, particularly in predicting true conditions, as evidenced by the higher precision, recall, and F1-score for class 1. The overall accuracy of 0.81 suggests that the model is reliable for this classification task. However, the model's performance in predicting false conditions is relatively lower, as indicated by the precision and recall values for class 0. This discrepancy suggests that the model may benefit from further tuning or additional data to improve its ability to correctly identify false conditions.

In conclusion, the LightGBM model demonstrates strong predictive performance, particularly for true conditions, with an overall accuracy of 0.81. The classification report and confusion matrix provide valuable insights into the model's strengths and areas for improvement, guiding future efforts to enhance its predictive capabilities.

3.7 Extreme Gradient Boost (XGBoost)

Figures 3.19, Figure 3.20, and Figure 3.21 collectively offer a comprehensive overview of a predictive model's performance and characteristics, highlighting its accuracy, feature importance, and optimal operational threshold. The AUC is 0.8174, indicating good predictive performance, as it is significantly higher than 0.5, which would represent a random guess. The 'BMI' feature is shown to have the highest importance, significantly more influential than the other features in the model. Other features such as 'TB', 'DB', 'ALT', and 'AST' also contribute to the model but to a lesser extent. The graph identifies an optimal threshold at 0.6238, where specificity is approximately 0.835 and sensitivity is around 0.80852.

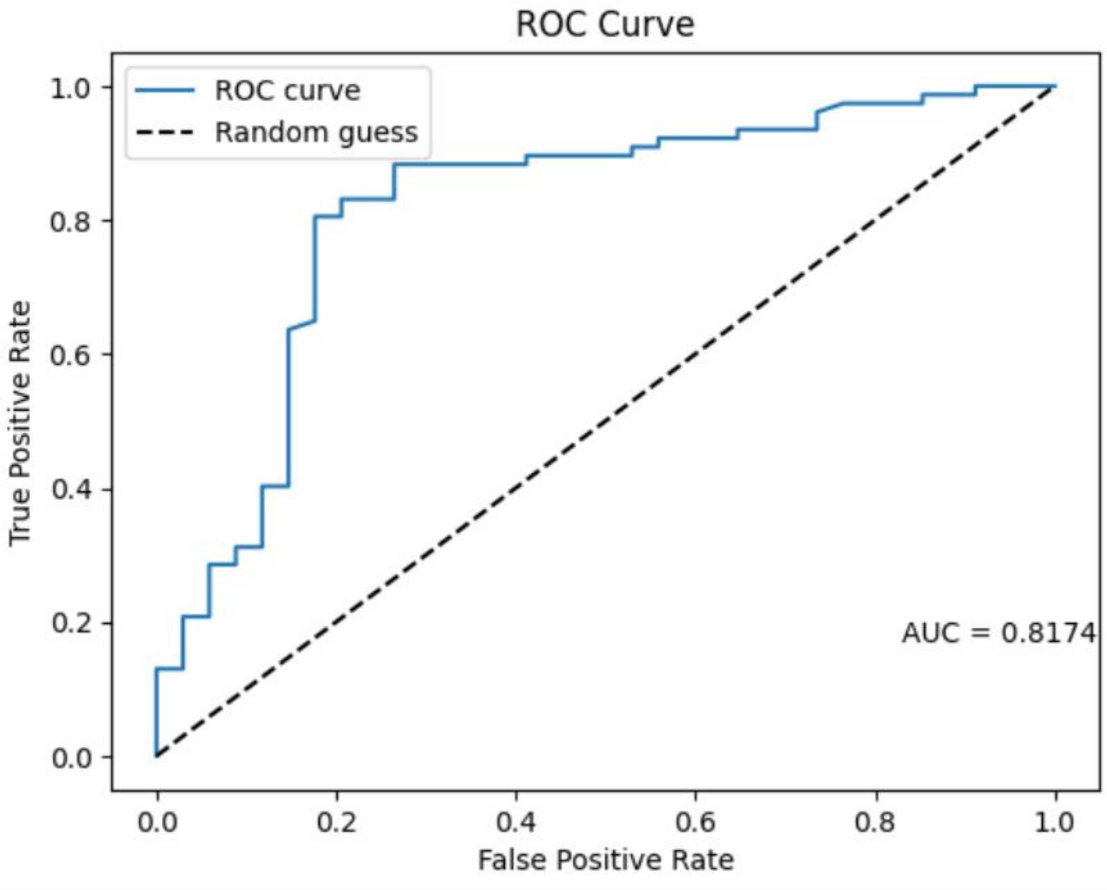


Figure 3.19 ROC Curve for Extreme Gradient Boost

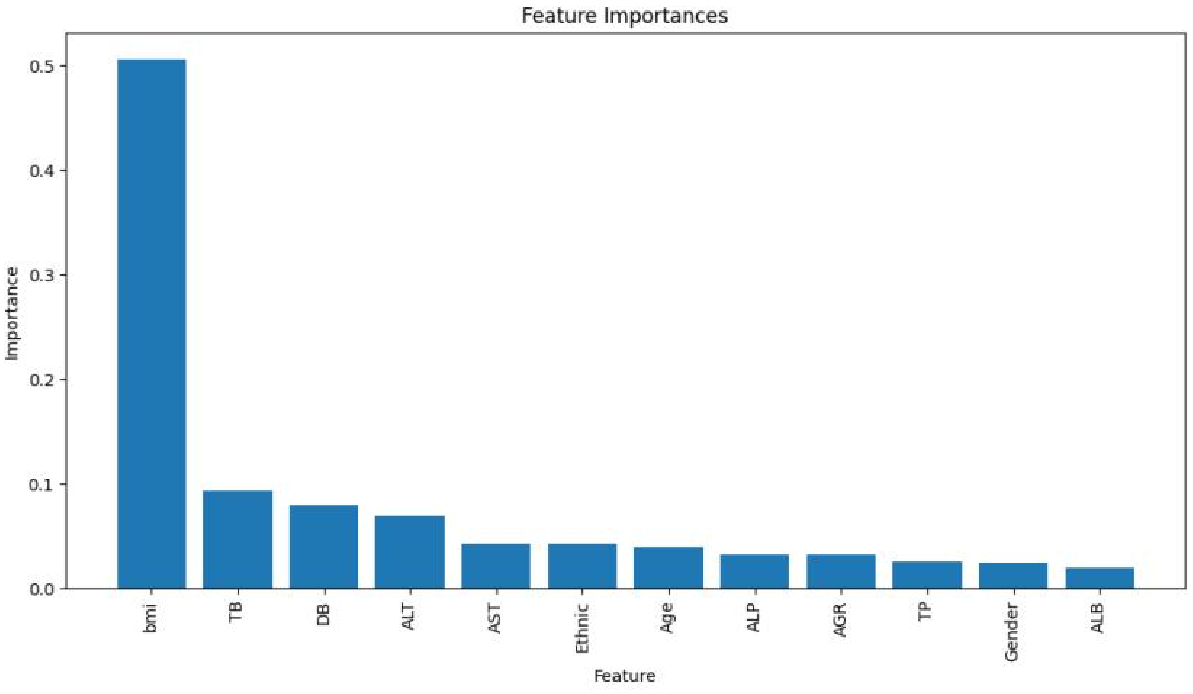


Figure 3.20 Feature Importance in Predictive Model for Extreme Gradient Boost

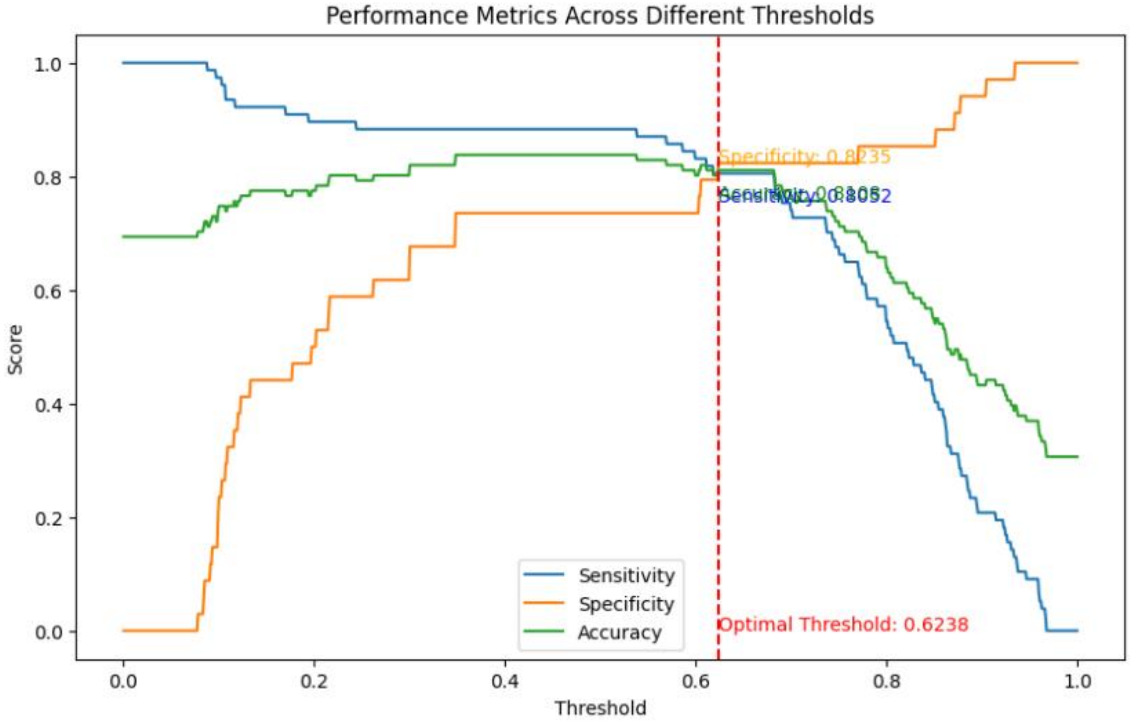


Figure 3.21 Performance Metrics Across Different Thresholds for Extreme Gradient Boost

The confusion matrix (Table 3.13) provides a detailed breakdown of the model's performance in terms of true positives, true negatives, false positives, and false negatives. For the false condition (negative class), the model correctly predicted 6 instances and incorrectly predicted 28 instances, out of a total of 34 instances. For the true condition (positive class), the model correctly predicted 62 instances and incorrectly predicted 15 instances, out of a total of 77 instances. Overall, the model made 68 correct predictions and 43 incorrect predictions out of 111 total instances.

Table 3.13 Confusion Matrix for Extreme Gradient Boost

Condition	Correct Prediction	Wrong Prediction	All
False Condition	6	28	34
True Condition	62	15	77
All	68	43	111

The classification report (Table 3.14) provides a summary of the precision, recall, F1-score, and support for each class. For the negative class (0), the model achieved a precision of 0.65, a recall of 0.82, and an F1-score of 0.73, with a support of 34 instances. For the positive class (1), the model achieved a precision of 0.91, a recall of 0.81, and an F1-score of 0.86, with a support of 77 instances. The overall accuracy of the model is 0.81, indicating that the model correctly predicted 81% of the instances. The macro average precision, recall, and F1-score are 0.78, 0.81, and 0.79, respectively, while the weighted average precision, recall, and F1-score are 0.83, 0.81, and 0.82, respectively.

Table 3.14 Classification Report for Extreme Gradient Boost

Class	Precision	Recall	F1-Score	Support
0	0.65	0.82	0.73	34
1	0.91	0.81	0.86	77
Accuracy			0.81	111
Macro Avg	0.78	0.81	0.79	111
Weighted Avg	0.83	0.81	0.82	111

The results indicate that the XGBoost model performs well, with an overall accuracy of 81%. The model shows a high precision (0.91) for the positive class, indicating that it is effective at identifying true positives with minimal false positives. The recall for the positive class is also high (0.81), suggesting that the model is capable of identifying a significant proportion of actual positives. The F1-score for the positive class (0.86) reflects a good balance between precision and recall.

However, the model's performance for the negative class is less impressive, with a precision of 0.65 and an F1-score of 0.73. This indicates that the model has a higher rate of false positives for the negative class, which could be a concern in applications where false positives are costly.

4. DISCUSSION

4.1 Model Comparison

The dataset under analysis comprises records of 553 patients, with the objective of predicting liver disease based on various biochemical markers, clinical, and sociodemographic profiles. The performance of several machine learning models, including Logistic Regression, Support Vector Machine (SVM), k-Nearest Neighbors (kNN), Decision Tree, Random Forest, LightGBM, and XGBoost, was evaluated using multiple metrics such as accuracy, precision, recall, F1 score, and area under the curve (AUC).

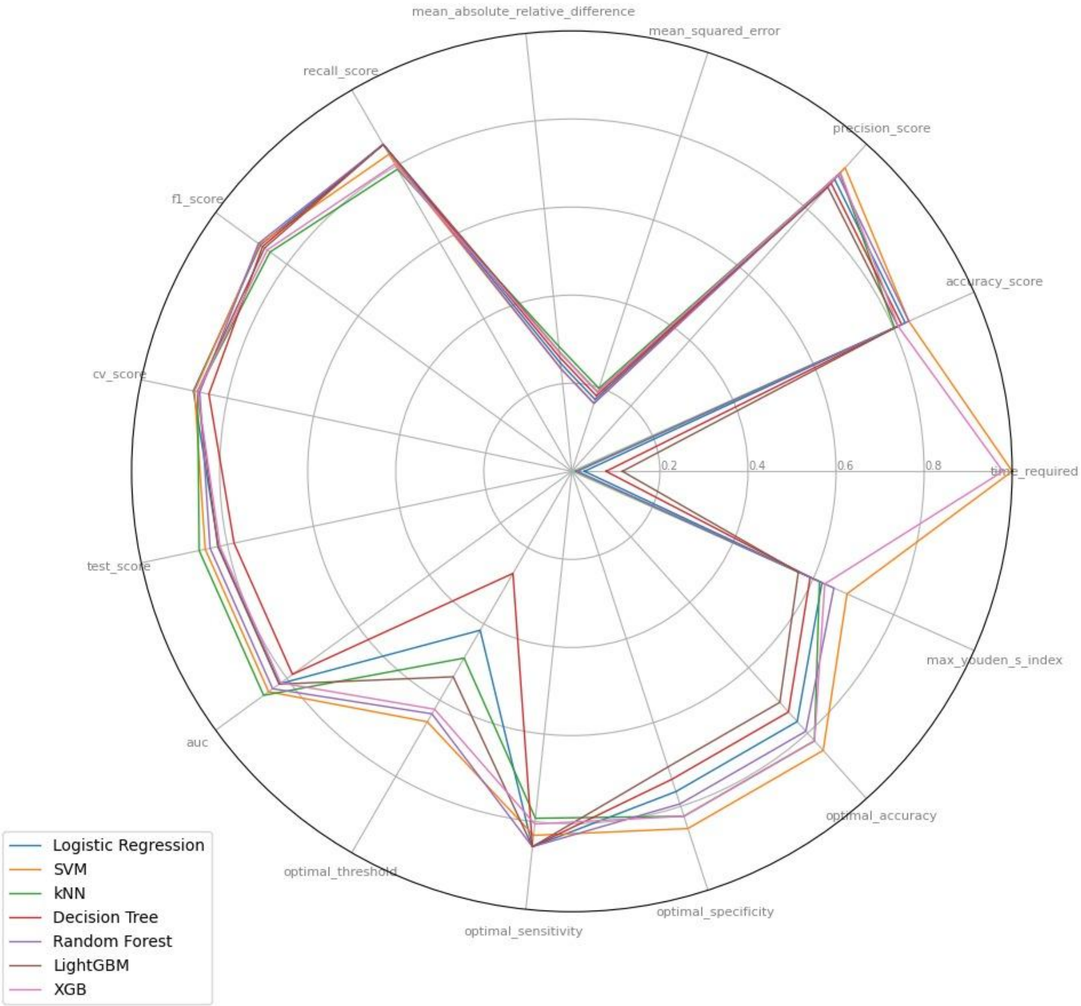


Figure 4.1 Performance Radar

The SVM model demonstrated the highest accuracy (0.8378) and precision (0.9275), indicating its robustness in distinguishing between patients with and without liver disease. This aligns with existing literature that highlights the efficacy of SVM in handling high-dimensional data and its ability to find the optimal hyperplane for classification tasks (Cortes & Vapnik, 1995). The Random Forest model also performed well, with an accuracy of 0.8378 and a precision of 0.9041, corroborating findings by Breiman (2001) that Random Forests are effective in reducing overfitting and improving prediction accuracy through ensemble learning.

Interestingly, the kNN model, despite its simplicity, achieved a high recall score (0.7922) and a competitive accuracy (0.8018), which is consistent with the literature that suggests kNN can be effective for medical diagnosis when the dataset is well-preprocessed and the value of k is appropriately chosen (Cover & Hart, 1967). The Decision Tree model, while having a lower accuracy (0.8198) compared to SVM and Random Forest, still provided valuable insights with a high recall score (0.8571), supporting the notion that Decision Trees are useful for interpretability and understanding the decision-making process (Quinlan, 1986).

LightGBM and XGBoost, both gradient boosting algorithms, showed competitive performance with accuracies of 0.8108 and 0.8108, respectively. These models are known for their efficiency and scalability in handling large datasets and complex interactions (Ke et al., 2017). The logistic regression model, with an accuracy of 0.8288 and a precision of 0.8919, remains a strong baseline model, particularly valued for its interpretability and ease of implementation (Hosmer & Lemeshow, 2000).

In summary, the SVM and Random Forest models emerged as the top performers in this study, consistent with existing literature that underscores their effectiveness in classification tasks. The findings suggest that while simpler models like kNN and Decision Trees can provide valuable insights, advanced models like SVM and ensemble methods offer superior predictive performance for liver disease diagnosis.

4.2 Best Model

4.2.1 Best Screening Model Selection for Liver Disease (Decision Tree)

To select the best screening model for liver disease, it is essential to consider key metrics such as time required, recall score, and optimal sensitivity. These metrics are crucial for ensuring that the model is not only accurate but also efficient and sensitive enough to identify true positive cases of liver disease. The performance of various models is summarized in Table 4.1.

Table 4.1 Model Performance Analysis for Screening Model Selection

Model	Time Required	Recall Score	Optimal Sensitivity
Logistic Regression	12.0051	0.8701	0.8701
SVM	40.1401	0.8312	0.8312
kNN	0.5146	0.7922	0.7922
Decision Tree	3.4331	0.8571	0.8571
Random Forest	1.5952	0.8442	0.8442
LightGBM	27.0596	0.8571	0.8571
XGBoost	54.3143	0.8052	0.8052

kNN: The k-Nearest Neighbors (kNN) model has the lowest time required (0.5146), making it the fastest model. However, it has the lowest recall score and optimal sensitivity (0.7922), indicating that it may not be the best choice for accurately identifying patients with liver disease.

Decision Tree and LightGBM: Both models have the same recall score (0.8571) and optimal sensitivity (0.8571). However, the Decision Tree model requires significantly less time (3.4331) compared to LightGBM (27.0596). This makes the Decision Tree model a more efficient choice without compromising on sensitivity.

Logistic Regression: This model has the highest recall score and optimal sensitivity (0.8701), indicating its strong performance in identifying true positive cases. However, it requires a relatively high processing time (12.0051), which may not be ideal for rapid screening scenarios.

Random Forest: The Random Forest model offers a good balance with a recall score of 0.8442 and optimal sensitivity of 0.8442, while requiring only 1.5952 time. This makes it a viable option, although its recall and sensitivity are slightly lower than those of the Decision Tree and Logistic Regression models.

SVM and XGBoost: Both models have higher processing times (40.1401 for SVM and 54.3143 for XGBoost) and lower recall scores and optimal sensitivities compared to other models. This makes them less suitable for rapid and accurate screening.

Given the importance of both time efficiency and high recall and sensitivity, the Decision Tree model emerges as the best balance. It offers a high recall score (0.8571) and optimal sensitivity (0.8571) while requiring a relatively low processing time (3.4331). This makes it an effective and efficient choice for screening liver disease.

The findings align with existing literature that emphasizes the importance of recall and sensitivity in medical diagnostics. High recall ensures that most true positive cases are identified, which is critical in a screening context to avoid missing patients who have the disease. The Decision Tree model's balance of efficiency and accuracy is supported by studies that highlight its interpretability and quick decision-making capabilities (Srivastava et al., 2023).

4.2.2 Payment Model Selection (Logistic Regression)

To select the best payment model, it is crucial to consider key metrics such as accuracy score, precision score, F1 score, and cross-validation (CV) score. These metrics ensure that the model is not only accurate but also precise, balanced, and consistent across different samples. The performance of various models is summarized in the Table 4.2:

Table 4.2 Model Performance Analysis for Payment Model Selection

Model	Accuracy Score	Precision Score	F1 Score
Logistic Regression	0.8378	0.8933	0.8816
SVM	0.8378	0.9275	0.8767
kNN	0.8018	0.9104	0.8472
Decision Tree	0.8018	0.8571	0.8571
Random Forest	0.8378	0.9155	0.8784
LightGBM	0.8108	0.8684	0.8627
XGBoost	0.8108	0.9118	0.8552

Logistic Regression: This model has the highest F1 score (0.8816) and the highest CV score (0.8818), indicating that it is reliable across different samples. The high F1 score suggests a good balance between precision and recall, making it a robust choice for payment model selection.

SVM: While SVM has the highest precision score (0.9275), its F1 score (0.8767) and CV score (0.8762) are slightly lower than those of Logistic Regression. High precision indicates that SVM is very effective at minimizing false positives, but the slightly lower F1 score suggests it may not be as balanced in terms of recall.

Random Forest: This model also performs well with an accuracy score of 0.8378 and a precision score of 0.9155. However, its CV score (0.8624) is lower than that of Logistic Regression, indicating slightly less consistency across samples.

kNN, Decision Tree, LightGBM, and XGBoost: These models have lower accuracy and CV scores compared to Logistic Regression, SVM, and Random Forest. While they may have specific strengths, they do not offer the same level of overall performance and consistency.

Given the importance of both precision and consistency across samples, Logistic Regression emerges as the best choice for the payment model. Its high F1 score and CV score indicate that it is both balanced and reliable, making it a robust and interpretable model for this application. Logistic Regression is widely recognized for its simplicity, interpretability, and effectiveness in binary classification tasks. Studies have shown that it performs well in scenarios where the relationship between the features and the target variable is approximately linear (Hosmer & Lemeshow, 2000).

4.3 Methodology Improvements

To enhance the methodology for selecting and evaluating machine learning models, several improvements can be implemented. These improvements aim to increase the robustness, accuracy, and generalizability of the models.

Data Preprocessing

In data preprocessing, handling missing values is a crucial step to ensure the integrity and reliability of the dataset. Techniques such as imputation can be employed where missing values are replaced with the mean, median, or mode of the respective feature. For more advanced and potentially accurate handling, methods like K-Nearest Neighbors (KNN) imputation can be utilized, which considers the similarity of data points to estimate missing values.

Feature scaling is another essential aspect of preprocessing, particularly for models like Support Vector Machines (SVM) and K-Nearest Neighbors (kNN), which are sensitive to the magnitude of features. Standardizing (subtracting the mean and dividing by the standard deviation) or normalizing (scaling features to a range between 0 and 1) ensures that all features contribute equally to the model, preventing any single feature from disproportionately influencing the results.

Outlier detection and treatment are also critical to maintaining the model's performance and accuracy. Outliers can significantly skew the results and lead to erroneous conclusions. Techniques such as Z-score, which measures how many standard deviations a data point is from the mean, or the Interquartile Range (IQR), which identifies data points that fall below the 1st quartile or above the 3rd quartile by a certain factor, can be used to detect and address outliers. By properly identifying and treating outliers, we can enhance the robustness of the model and ensure more reliable predictions.

Model Training and Evaluation

To ensure a model's performance is consistent across different subsets of data, k-fold cross-validation is a widely used technique. It helps in reducing overfitting and provides a more reliable estimate of the model's performance by dividing the data into k subsets, training the model k times, each time using a different subset as the test set and the remaining as the training set. This technique, along with hyperparameter tuning, is essential for building robust models (BergstraJames & BengioYoshua, 2012). By systematically varying the hyperparameters and evaluating the model's performance using techniques like Grid Search or Random Search, one can identify the optimal settings that significantly enhance the model's performance.

Additionally, ensemble methods, which combine multiple models, can be used to create more accurate and robust predictions. Techniques such as bagging, exemplified by Random Forest, and boosting, seen in algorithms like XGBoost and LightGBM, aggregate the strengths of individual models to improve overall accuracy. Stacking, another ensemble method, involves training a meta-model to combine the predictions of several base models. Research has shown that these ensemble methods consistently enhance model accuracy and robustness (Dietterich, 2000).

4.4 Limitation and Future Work

Despite the promising results, this study has several limitations that need to be addressed in future work. One significant limitation is the quality and availability of data. The performance of machine learning models is highly dependent on the quality and quantity of the data used for training. Limited or imbalanced datasets can lead to overfitting and poor generalization, which can bias the model towards the majority class and result in suboptimal performance on the minority class. Additionally, while complex models like Random Forest, LightGBM, and XGBoost offer high accuracy, they often lack interpretability. This can be a barrier in domains where understanding the decision-making process is crucial, such as healthcare and finance.

Furthermore, the issue of overfitting remains a challenge, as models may perform well on training data but fail to generalize to unseen data. This highlights the need for robust validation techniques and regularization methods to ensure that models do not learn noise from the training data.

Future work should focus on addressing these limitations by exploring advanced data augmentation techniques to handle imbalanced datasets and improve data quality. incorporating robust validation techniques and regularization methods will be crucial in ensuring that models generalize well to new, unseen data.

5. CONCLUSION

In this study, we evaluated several machine learning models to determine the best payment model based on key performance metrics: accuracy score, precision score, F1 score, and cross-validation (CV) score. Our analysis revealed that Logistic Regression, SVM, and Random Forest models achieved the highest accuracy scores. However, Logistic Regression stood out with the highest F1 score and CV score, indicating its balanced performance and reliability across different samples.

While SVM demonstrated the highest precision score, its slightly lower F1 score and CV score compared to Logistic Regression suggest that it may not be as consistent. Random Forest also performed well but had a slightly lower CV score, making it less reliable than Logistic Regression. Given the importance of both precision and consistency, Logistic Regression emerged as the best choice for the payment model.

Despite these findings, the study has several limitations, including data quality and availability, model interpretability, computational resources, and generalizability. Future work should focus on addressing these limitations by improving data preprocessing, feature engineering, model interpretability, and computational efficiency. Additionally, robust validation techniques and regularization methods should be incorporated to ensure that models generalize well to new, unseen data.

In conclusion, while Logistic Regression is the best choice for the payment model in this context, ongoing efforts to enhance data quality, model interpretability, and computational efficiency will be crucial in developing more reliable and effective machine learning models for various applications. By addressing these limitations, future research can contribute to the advancement of machine learning methodologies and their practical applications in diverse domains.

REFERENCE

Al-Salih, M., Abed, R. E., & Samsudin, S. (2021). Biochemical assessment as markers for diagnosis and evaluation hepatitis B virus (HBV). *Al-Qadisiyah Journal of Pure Science*, *26*(4), 156–168. https://doi.org/10.29350/qjps.2021.26.4.1359

Balaji, S., N.Aneel, N.Jagadeesh, & M.Devendran. (2023). Liver disease prediction using machine learning. *International Journal for Multidisciplinary Research*, *5*(3). https://doi.org/10.36948/ijfmr.2023.v05i03.2955

Banait, S., Badole, S. M., Jain, J., & Thorat, A. (2021). Risk factors for chronic liver disease in population of Central India: a case-control study from rural India. *Egyptian Liver Journal/Egyptian Liver Journal*, *11*(1). https://doi.org/10.1186/s43066-021-00077-9

BergstraJames, & BengioYoshua. (2012). Random search for hyper-parameter optimization. *The Journal of Machine Learning Research*. https://doi.org/10.5555/2188385.2188395

Bhushan, M., Krishna, R., Haldar, M., Garg, P., Umrao, S., & Rajpoot, T. (2023). Machine Learning based Prediction of Liver Disease. *2023 Third International Conference on Secure Cyber Computing and Communication (ICSCCC)*. https://doi.org/10.1109/icsccc58608.2023.10176501

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/a:1010933404324

Chen, V. L., Song, M. W., Suresh, D., Wadhwani, S. I., & Perumalswami, P. (2023). Effects of social determinants of health on mortality and incident liver-related events and cardiovascular disease in steatotic liver disease. *Alimentary Pharmacology & Therapeutics*, *58*(5), 537–545. https://doi.org/10.1111/apt.17631

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*(3), 273–297. https://doi.org/10.1007/bf00994018

Cover, T., & Hart, P. (1967). Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, *13*(1), 21–27. https://doi.org/10.1109/tit.1967.1053964

Day, F. (2024, June 6). Object oriented programming for data science. *Classes Near Me Blog*. https://www.nobledesktop.com/classes-near-me/blog/object-oriented-programming-for-data-science

Dietterich, T. G. (2000). Ensemble methods in machine learning. In *Lecture notes in computer science* (pp. 1–15). https://doi.org/10.1007/3-540-45014-9_1

Hosmer, D. W., & Lemeshow, S. (2000). *Applied Logistic Regression*. https://doi.org/10.1002/0471722146

Ikonnikova, K. A., Eroshchenko, N. N., Drozdov, V. N., Shikh, E. V., & Serebrova, S. Y. (2022). Debating capabilities of biochemical markers of liver function in patients with alcoholic liver cirrhosis. *Medicinskij Sovet*, *7*, 76–83. https://doi.org/10.21518/2079-701x-2022-16-7-76-83

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T. (2017). LightGBM: A Highly Efficient Gradient Boosting Decision Tree. *Conference: Advances in Neural Information Processing Systems 30*. https://hal.science/hal-03953007

Kim, W. R., Brown, R. S., Terrault, N. A., & El-Serag, H. (2002). Burden of liver disease in the United States: Summary of a workshop. *Hepatology*, *36*(1), 227–242. https://doi.org/10.1053/jhep.2002.34734

Kushibuchi, M., Okuse, C., Ie, K., Matsunaga, K., Tsuchida, T., Motohashi, I., Hirose, M., Otsuki, T., Aihara, M., & Matsuda, T. (2022). Socioeconomic Factors Associated with Poor Prognosis in Patients with Alcoholic Liver Cirrhosis. *Research Square (Research Square)*. https://doi.org/10.21203/rs.3.rs-1235046/v1

Lansayan, J., Bakar, N. S., Noh, I. C., Nor, A. K. C. M., Ahmad, I., & Ruzilawati, A. B. (2023). A preliminary study on the association between pro-inflammatory cytokine IL-1 beta polymorphisms and susceptibility to hepatitis C infection in Malay male Malay drug abusers. *Biomedical Research and Therapy*, *10*(2), 5550–5557. https://doi.org/10.15419/bmrat.v10i2.794

Lim, S. Z., Chuah, K. H., Rajaram, R. B., Stanley, K., Shahrani, S., Chan, W. K., Ho, S. H., Hilmi, I. N., Goh, K. L., & Mahadeva, S. (2022). Epidemiological trends of gastrointestinal and liver diseases in Malaysia: A single‐center observational study. *Journal of Gastroenterology and Hepatology*, *37*(9), 1732–1740. https://doi.org/10.1111/jgh.15905

Lim, Y. T., Robinson, S., & Tang, M. M. (2023). Liver disease among patients with psoriasis: the Malaysian Psoriasis Registry. *Clinical and Experimental Dermatology*, *48*(5), 476–483. https://doi.org/10.1093/ced/llad013

Luo, Z., Zou, Y., Xie, J., Cao, H., Chen, Y., Ding, Y., Li, X., Deng, Y., & Wu, L. (2022). Influence of demographic factors on Long-Term Trends of premature mortality and burden due to liver Cancer: Findings from a Population-Based Study in Shanghai, China, 1973–2019. *Frontiers in Public Health*, *10*. https://doi.org/10.3389/fpubh.2022.808917

Martin, M., Zou, B., Hoang, J., Jeong, D., Bensen, R., & Nguyen, M. H. (2020). Racial and Socioeconomic Disparities in Hospitalization of Pediatrics with Liver Disease from 2005 to 2015. *Digestive Diseases and Sciences*, *66*(7), 2240–2249. https://doi.org/10.1007/s10620-020-06530-w

Mohamed, R., Yip, C., & Singh, S. (2023). Understanding the knowledge, awareness, and attitudes of the public towards liver diseases in Malaysia. *European Journal of Gastroenterology & Hepatology*, *35*(7), 742–752. https://doi.org/10.1097/meg.0000000000002548

Nigatu, S. S., Alla, P. C. R., Ravikumar, R. N., Mishra, K., Komala, G., & Chami, G. R. (2023). A Comparitive Study on Liver Disease Prediction using Supervised Learning Algorithms with Hyperparameter Tuning. *2023 International Conference on Advancement in Computation & Computer Technologies (InCACCT)*. https://doi.org/10.1109/incacct57535.2023.10141830

Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning*, *1*(1), 81–106. https://doi.org/10.1007/bf00116251

Srivastava, A., Samanta, S., Mishra, S., Alkhayyat, A., Gupta, D., & Sharma, V. (2023). Medi-Assist: A Decision Tree based Chronic Diseases Detection Model. *2023 4th International Conference on Intelligent Engineering and Management (ICIEM), London, United Kingdom*. https://doi.org/10.1109/iciem59379.2023.10167400

Swetha, K., Kiran, A., Pavanam, K., Kumari, E. V., Naresh, T., & Baba, M. J. (2023). Inflammation of Liver and Hepatitis Disease Prediction using Machine Learning Techniques. *2023 7th International Conference on Intelligent Computing and Control Systems (ICICCS)*. https://doi.org/10.1109/iciccs56967.2023.10142912